MACHINE LEARNING AND BIG DATA ANALYTICS FOR THE SMART GRID

A Dissertation
Presented to
The Academic Faculty

by

Xiaochen Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2017

COPYRIGHT © 2017 BY XIAOCHEN ZHANG

MACHINE LEARNING AND BIG DATA ANALYTICS FOR THE SMART GRID

Approved by:

Dr. Santiago Grijalva, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Lukas Graber
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Ronald Harley
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Duenhorng Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Maryam Saeedifard
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: July 24, 2017

To my parents, Huigang and Xiaoming,

with all my love.

ACKNOWLEDGEMENTS

The process of fulfilling my Ph.D. degree at Georgia Tech is one of the most important experiences in my life that I will always treasure. During the past four years, I have received so much help and support from my family and friends, my fellow lab mates, and professors.

First, I would like to thank my parents, Huigang and Xiaoming, for their support and encouragement that made me who I am today.

I would like to thank my advisor, Dr. Santiago Grijalva, who is supportive, generous, and is always there to provide guidance to his students. Since my first day in the advanced computational electricity systems (ACES) lab, Dr. Grijalva gave me the full freedom to follow my own interests. It is not only his guidance on my research challenges, but also his passion and positive attitude that inspired me and carried me through the ups and downs of my student life. Later on, I was also given an opportunity to pursue a second degree in data science, which is the foundation of this research work.

I also owe special thanks to Dr. Ronald Harley, Dr. Maryam Saeedifard, Dr. Lukas Graber, and Dr. Polo Chau for serving as my dissertation committee. All of them provided invaluable comments and guidance on my dissertation.

I would like to express my gratitude to the professors, who inspired and guided me through the various stages of my research. Thanks to Dr. Yajun Mei and Dr. Benjamin Haaland who guided me on photovoltaic system detection research. Thanks to Dr. Sigrún Andradóttir who helped me on electrical vehicle charging behavior research. Thanks to Dr.

Polo Chau who provided me with great computational tools for the time variant load modeling task.

Finally, I would like to thank my ACES lab mates. I am honored to have worked with all these talented minds: Jouni Peppanen, Matthew Reno, Tanguy Hubert, Leilei Xiong, James Thomas, Jose Grimaldo, and Jeremiah Deboever, all of whom helped me on my academic and non-academic matters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

As numerous sensors, such as smart meters and PMUs, continue to be added to the grid, the emerging information collected is becoming a valuable source to researchers and grid operators who seek to conduct advanced analytics on the smart grid. This research combines the latest machine learning and big data analytics techniques with the domain knowledge of the smart grid to explore the added value of the emerging power system data. By exploiting the emerging smart grid database, we can develop data-driven solutions for the most pressing issues, such as load modeling, demand side management, and distributed energy resource hosting capacity analysis. This research first develops a methodology to apply data science technologies to smart grid applications. Then, it provides a set of examples to illustrate how the smart grid may benefit from the emerging data. These examples cover a broad range of smart grid analyses and applications, including residential photovoltaic system detection, electrical vehicle charging demand modeling, time-variant load modeling, and hosting capacity analysis. Different data analytics techniques are implemented in these examples, including clustering, statistical inference, change-point detection, parameter estimation, stochastic modeling, and statistical learning methods.

CHAPTER 1. INTRODUCTION

1.1 The Evolution of the Electric Power Grid

The continuous evolution of the power grid, which includes generation, transmission, and distribution, is rapidly changing the smart grid, and impacting its planning and operation. The current power grid is rapidly growing in complexity due to three fundamental forces: new devices, new data sources, and more demanding social and environmental requirements. These three driving forces interact with each other and push the development of the power grid in the fields of technology, markets, and public policy.

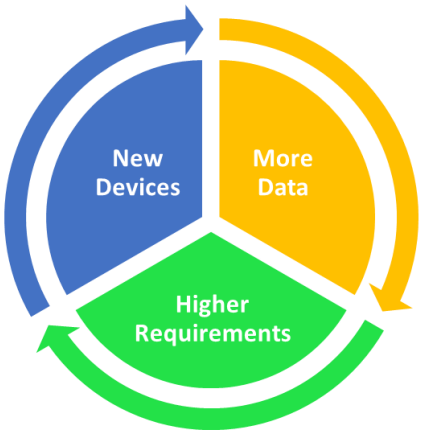

Figure 1. Three driving forces of the future smart grid.

The social and environmental goals of cleaner energy and the awareness of global warming have led to a rapid deployment of new devices and services including wind, solar, demand response, and electric vehicles. According to the U.S. Environmental Protection Agency (EPA), by 2040, renewables will surpass coal and nuclear as the second electricity generation source next to natural gas [1] [2]. According to the U.S. Energy Information

Administration (EIA), wind and solar will become the major source of growth for global renewable energy generation for decades to come [3]. The transportation system is the second largest energy consumption sector next to electricity in the U.S. As the power system continues to provide cleaner electricity, electrical vehicles (EVs) become the choice for commuting adopted by the general public, and are encouraged by local government as well. For the last five years, the global annual sale of light-duty EVs has experienced an almost exponential growth, which dramatically changes traditional electricity demand [4].

The higher penetration of the renewables in the grid as well as the fast adoption of demand response and electrical vehicles, in turn, calls for advanced grid control strategies to avoid potential side effects. These potential problems include power quality issues, system reliability issues, and the financial sustainability of the utility business model. Most of the solutions for these pressing issues are supported by rich data collected through new metering devices such as phasor measurement units (PMUs), intelligent electronic devices (IEDs), and smart meters. The emerging data accumulated at all levels of the smart grid enable researchers and power system operators to advance the power system to operate in a more secure, economic and sustainable manner.

1.2 Machine Learning and Data Analytics for Smart Grid

New data sources such as PMUs, micro PMUs, intelligent electronic devices (IEDs), and smart meters are becoming standard in modern power grids. According to a report from National Renewable Energy Laboratory (NREL) [5], a typical synchronized PMU calculates and stores 30-60 data points per second. A network with 5 PMUs can generate a 13.8MB report every hour. Residential smart meters may generate data at a much lower

rate; however, smart meters are much more common, thus are generating data of comparable size. According to the Edison Foundation [6], with more than 40 percent of the U.S. households having a smart meter, more than 46 million smart meters have been installed in the U.S. by 2015, which daily generate 1 billion data points.

Traditional power systems have limited number of sensors, thus data are collected at an aggregated level. These data are usually of small size and with a lower sampling frequency. When dealing with data of small size, human expertise along with simple statistical analysis and regression is enough for most applications such as load forecasting. For example, utilities are satisfied with an aggregated load model when only hourly energy consumption at the substation level is available. However, today's smart grid collects much larger volumes of data. These data are collected closer to end use and at a much higher frequency. This results in new data sets that are fast growing in size and complexity. Due to the size of the data and the ubiquitous correlations among different data sources, machine learning and data analytics techniques are the logical solutions to exploiting the otherwise hidden value of the smart grid data. First, it is no longer possible to visually inspect and analyze potential abnormalities in this huge data set through human labor. In addition, important power system behaviors and patterns are usually hidden under the substantial noise and randomness of the measurements that are collected closer to end use. This adds another layer of difficulty for human experts to manually discover insights or statistical patterns from the newly available smart meter data.

Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [7]. Machine learning seeks to provide increasing levels of automation in the knowledge engineering process, replacing

much time consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data [8]. Thus, there is a huge potential for implementation of the latest machine learning and big data analytics technology to the evolving smart grid.

Machine learning and artificial intelligence have a long history of applications in power system control and analysis. In early 1990s, machine learning was first introduced in power systems to perform system fault diagnosis, including fault detection and fault classification [9]. However, due to the computational capability of machines at that time being, most of the machine learning algorithms were only used as extra supports to human experts.

As the continuous development of information technology, machine learning has become an independent subject apart from artificial intelligence (AI), and has started to flourish. Instead of focusing on neural network and traditional signal processing algorithms such as wavelet analysis, researchers started to explore alternatives such as fuzzy logic and support vector machines. In early 2000s, due to the deregulation of power systems and the development of electricity markets, more machine learning algorithms were implemented on applications such as load forecast, network reconfiguration, energy price prediction, and trading strategies. More recently, many new issues have raised in the power industry by the accelerated adoption of renewable energy. Renewable energy output prediction becomes a suitable application of machine learning and data analytics.

Apart from solving the traditional power grid problems, the latest machine learning technology provides additional services to the grid. For example, utilities and system

operators may learn the consumer's energy consumption patterns through mining the smart meter measurements. Advanced video processing algorithms can be used for substation theft detection, which improves system security and reduces operational costs. Advanced image processing and pattern recognition may even allow machines to take over some basic and routine workload from human labor.

In summary, machine learning and data analytics have great potential to deepen researchers' understanding of the power grid, to provide power system planners with additional degrees of freedom to build a greener system, and to allow system operators to operate the grid in an optimal state.

1.3 Research Objectives

This research aims at combining the latest machine learning and big data analytics techniques with the smart grid domain knowledge to explore the added value of the emerging power system data. The objectives of this research are listed below.

1. Explain the importance and opportunity of machine learning and data analytics in the planning and daily operation of the power grid. .
2. Provide a basic framework for power system data analytics.
3. Demonstrate the applications of machine learning and data analytics in four smart grid tasks.
    a) Residential photovoltaic system detection
    b) Electrical vehicle charging demand modeling
    c) Time-variant load modeling
    d) Hosting capacity analysis

1.4 Frameworks for Power System Data Analysis

The integration of the latest data analytics technologies and smart grid databases creates a promising opportunity for enhancing smart grid operations. Most of the smart grid data analytics follow a six-step process as shown in Figure 2.

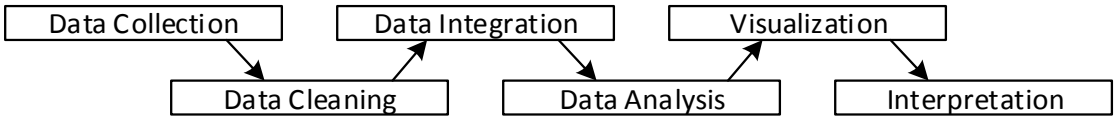

Figure 2. Power system data analytics process.

Smart grid data analytics always starts with data collection and cleaning. Data cleaning is a task of eliminating abnormalities of data sets. Some common data-related issues in smart grid measurements include missing data, wrong measurements, and non-synchronized time stamps. Most of these abnormalities can be detected using common physical constraints of the grid and statistical tests. Missing measurements can be filled using historical measurements and statistical methods such as multiple imputation.

Modern power systems accumulate massive data and information from many sources such as weather/geographic database, smart meter/PMU measurements, and market signals, as shown in Figure 3. Thus, it is important to integrate information from various sources and exploit their potential correlations.

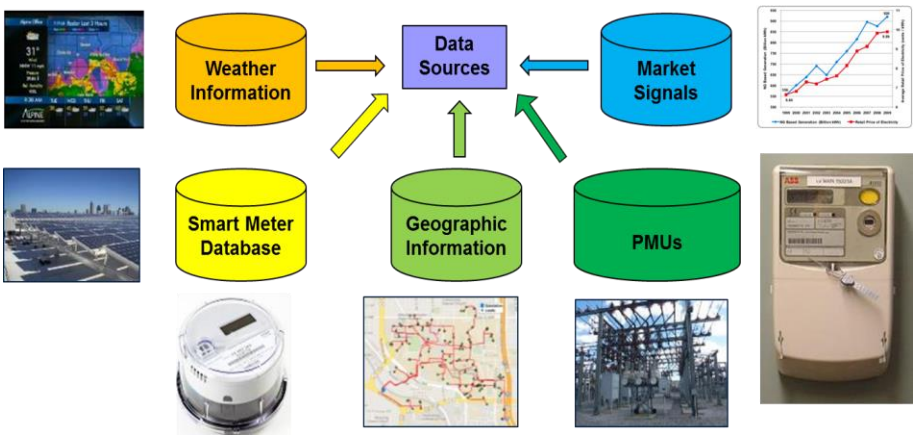

Figure 3. Power system data sources.

Data collection, cleaning, and integration lay a solid foundation for data analytics. Machine learning and statistical learning are the most common data analytics methods, which can be used in power system applications. For example, clustering algorithms can be used to create load profiles that cluster consumer behaviors based on their similarities; classification can be used for automatic fault diagnostics; and regression can be used for load forecasting and electricity trading strategy analysis.

The last step of smart grid data analytics is visualization and interpretation. Data visualization is commonly used in new data sets exploration and system fault diagnosis. Visualization is not necessary if the application does not involve human inputs. However, for any decision-making process that involves human inputs, visualization is the key to facilitate human-machine interactions. The right visualization tool can significantly shorten the decision-making process for power system operators.

1.5 Outline of Chapters

In Chapter 2, a data-driven solution is provided for unauthorized residential PV system detection. The proposed solution utilizes smart meter measurements and local

weather information as the major data sources. The method follows a three-step process. First, because any unauthorized PV system installation will cause a change in energy consumption patterns, a change-point detection algorithm is used to screen abnormalities in consumer behaviors. In the second step, statistical inference is constructed to verify whether the detected abnormality is caused by an unauthorized PV system. In the third step, the parameters of the PV system are estimated using the smart meter measurements and local whether information.

In Chapter 3, a stochastic model is developed to describe the charging demand of residential electrical vehicles. The proposed model captures the non-homogeneity and periodicity of the residential EV charging behavior through a self-service queue with a periodic and non-homogeneous Poisson arrival rate, an empirical distribution for charging duration, and a finite calling population. We validate the model by comparing the simulated time-series data with real measurements. The hypothesis test shows the proposed model accurately captures the EV charging behavior.

In Chapter 4, a novel time-variant load modeling method is proposed through mining of smart meter historical data. Given the data resolution (15 minutes per reading) in the database, the load's P-V and Q-V properties are extrapolated through clustering and regression. The historical measurements of the load provide the opportunity for creating a time-variant load model, which can be estimated using measurements at different time periods. The new load modeling method belongs to neither the component-based approach nor the measurement-based approach, and it is demonstrated using the real measurements collected from the Georgia Tech campus testbed.

In Chapter 5, a machine-aided hosting capacity analysis method is proposed to enable the safe interconnection of renewable energy on distribution networks. We first discuss advanced tools for hosting capacity analysis including the quasi static time-series (QSTS) simulation. We describe the limitations of QSTS associated with its computational time requirements. The major contribution is speeding up the QSTS simulation so that fast and accurate hosting capacity analysis can be achieved. The fast QSTS simulation task is formulated as a machine learning problem. We first discuss the limitations of using pre-existing machine learning methods. Then we propose a machine-aided method which feeds the machine learning black box with some knowledge of the physical distribution network. The proposed method achieves high computational time reduction for hosting capacity analysis with accurate results.

Chapter 6 summarizes the key results and concludes the research. Future research opportunities and potential technology development directions are also discussed.

CHAPTER 2. PHOTOVOLTAIC SYSTEMS DETECTION

A variety of distributed energy resources (DER) such as solar photovoltaic (PV) systems, micro turbines, and electrical vehicles (EV) are being connected to the grid [10]. According to the Hawaiian Electric Company (HECO), in 2015 one in eight of HECO's 450,000 customers has a residential PV system. As the speed of residential PV adoption continues to accelerate, in a high PV penetrative environment, utilities are facing technical problems related to overvoltage, frequency control, and back feeding flow [11], as well as financial issues such as a rapid decrease in revenue. In order to manage these new challenges, it is critical for utilities to gain visibility of all plugged-in PV systems, especially at the residential level.

2.1 Reliability and Financial Risks of Unauthorized PV Systems

Unauthorized PV installations may create safety risks, and lack of visibility may result in incorrect planning and operation, which leads to over-voltages, back-feeding, and in the worst-case scenario, damaging system equipment such as transformers, voltage regulators, and customers' appliances [12]-[13]. In order to facilitate the adoption of residential PV systems and to minimize risks, utilities enforce regulations and permits for residential PV systems. In California, Hawaii, and other states, it is required by law [14], [15] that customers should obtain necessary permits from permit agencies before any PV system installation. According to the DOE's report on smart grids 2014 [16], massive adoption of PV systems will lower the utilities' revenue, which in return increases the electric rate for non-solar customers. In Arizona, a fixed charge for new customers who

sign a contract with a solar energy provider was recently implemented [17], which leads to similar debates about solar interconnection fees in many states.

There are various reasons for unauthorized or incorrectly registered PV systems:

a) Owner decided not to apply for a permit to avoid permit fees [18],
b) Regulations were required after the system was installed,
c) Lack of awareness by the owner of diverse permitting rules by country, state, city, and often zonal regulations,
d) Different rules depending on the size and type of PV installation can make the owners believe they do not need a permit,
e) Changes in property ownership including transfers,
f) Multiple systems installed or future additions at the same premises,
g) Incorrect third party handling of the permit application, and
h) Data entry and data maintenance errors.

In 2014 Hawaii, the system with the highest penetration of PV in the US, recognized a large number of unauthorized PV installations [12] and prompted a specific program designated to reduce the number of these systems. In North Belgium, the number of unauthorized PV systems has exceeded that of the PV systems installed under the local certificate due to the introduction of the grid fee in 2013 [18]. This creates a serious problem for the operation and long-term planning of distribution systems.

An effective and efficient PV system detection and estimation algorithm can be proven to be of significant value to utilities for safety, reliability, and revenue reasons [19]. If not accurately modeled and managed, the fast adoption of PV systems in the distribution

system can put system security and reliability at risk. Traditional distribution networks are designed for one-directional power flow. High penetration of PV can lead to reverse power flow along the distribution feeders [20] and cause system protection failures. In addition, PV output is heavily influenced by sky cloud cover and can be highly variable, resulting in numerous energy spikes, transient over-voltage [13], and increased transformer tap-change operations.

Many researchers have studied the impacts and risks of PV on distribution systems [21]-[25], however, the detection and monitoring of residential PV systems has not been the focus of the studies and related research. As a result, we propose a data-driven approach to detect and monitor unauthorized and misfiled PV systems by implementing advanced data-mining algorithms on the smart meter data stream [26].

2.2 A Change-point-detection-based Solution

Thanks to the significant investment in smart meters and advanced metering infrastructures (AMI) in the past few years, the database populated with smart meter measurements is starting to play a very important role in utilities' daily operations, such as enhanced load forecasting [27], load modeling, demand response, and load profiling [28]-[29]. In this section, we show that the historical data collected by smart meters can help utilities detect unauthorized or misfiled PV systems in order to enhance their customer models. Better models and accurate databases result in significant operational benefits to the utility. The proposed method consists of three steps:

- Step 1: Unauthorized PV system screening
- Step 2: PV system verification test

- Step 3: PV size estimation.

In the first step, we detect energy consumption abnormalities among all customers using a recently developed change-point detection algorithm [30], which returns the abnormalities as change-points in the energy consumption time-series data. In the second step, we estimate the typical load profiles (TLP) before and after the change point using Gaussian kernel density estimation, which filters out noises that result from the customer's random behaviors. We construct a statistical inference using the permutation test with Spearman's rank coefficient to verify whether the change-point is caused by an unauthorized PV installation. Finally, once an unauthorized PV installation has been confirmed by statistical inference, we further estimate the size (rated power) of the detected PV system using the local cloud cover index (CCI). CCI is a numerical measure of the fraction of the sky obscured by clouds [31]. The proposed method has been validated on realistic system data sets, where all load components including PV outputs are recorded through separate meters.

2.3 Unauthorized PV System Detection

In this section, we discuss the organization of the smart meter measurements used in this study and formulate the residential PV detection problem as a combination of a change-point detection problem and a statistical inference.

2.3.1 Problem Formulation

2.3.1.1 Smart Meter Time Series Data

The data set used in this study corresponds to a set of 15minutes-resolution smart meter readings from hundreds of homeowners from a U.S. city, in 2013. Around 40 of these homeowners have home solar systems installed, and the corresponding 15 minutes-resolution PV outputs for each house are recorded through separate meters. The energy consumption and PV output data from these 40 PV-equipped houses are used in our study.

We model the smart meter historical readings as time-series streamed data, with the frequency of f readings per day ($f = 96$ in this study). Let $y(t_{d,i})$ denote the ith reading for day d, and let

$$\boldsymbol{D}(d) \coloneqq \begin{bmatrix} y(t_{d,1}) & y(t_{d,2}) & \cdots & y(t_{d,f}) \end{bmatrix}^T \in \mathbb{R}^f \tag{1}$$

denote the sequence of smart meter readings for day d. We batch the daily measurements into a data bundle $\boldsymbol{Y}(d)$ as in equation (2), where the time window is k days. Then,

$$\boldsymbol{Y}(d) := [\boldsymbol{D}(d) \quad \boldsymbol{D}(d+1) \cdots \boldsymbol{D}(d+k-1)] \in \mathbb{R}^{f \times k} \tag{2}$$

corresponds to all the smart meter readings starting from day d to day $(d + k - 1)$. The data bundle $\boldsymbol{Y}(d)$ is later used as input for the change-point detection algorithm. The data structure is further illustrated in Figure 4, where the change-point detection algorithm compares the differences between every two adjacent data bundles.

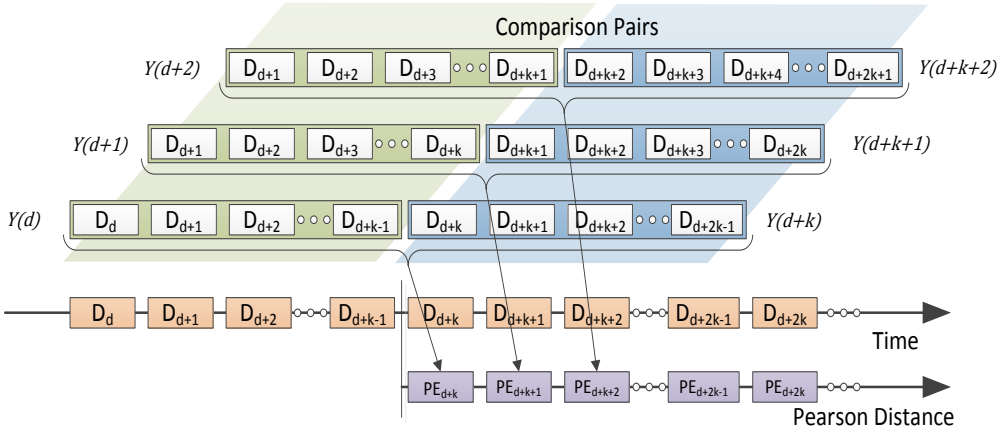

Figure 4. Time series data formulation

2.3.1.2 PV Detection Problem Formulation

The residential PV system installation detection can be formulated as a change-point detection problem. Let us consider a PV system installed at day $(d + k)$. The PV energy output will be reflected on the smart meter measurements of the customer. As a result, the smart meter readings or the data bundles before and after the PV installation date (e.g., $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d + k)$) must be dissimilar. We use Pearson divergence (PE divergence) to measure the dissimilarity between two different data bundles $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d + k)$, see equation (2) [32]. The change-point is detected based on the PE divergence score tested on every adjacent pair of the data bundles, as shown in Figure 4. Let us assume P and P' to be the distribution of the data in data bundles $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d + k)$, then $PE(P||P')$ is the PE divergence between distribution P and P', which can be computed using (3).

$$PE(P||P') \coloneqq \frac{1}{2}\int p'(\boldsymbol{Y})\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - 1\right)^2 d\boldsymbol{Y} \tag{3}$$

where $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are the probability density functions of the two distributions P and P'.

2.3.2 Relative Density-Ratio Estimation

Change-point detection or change-point analysis is a powerful tool used to detect abrupt changes in time series data. This method has been widely applied in many areas such as climate change [33], image processing [34] and financial economics [35]. Most change-point detection methods can be categorized into two classes: real-time detection and off-line detection. We adopt a recently developed off-line detection method that uses relative density-ratio estimation to detect abnormalities in customer energy consumption [30]. For a time-series data set, the change-point detection algorithm can detect various changes, such as jumping mean, scaling variance, switching covariance, or even varying frequency caused by PV installation.

The change-point detection algorithm developed in [30] is used here due to its efficiency and non-parametric nature. In equation (3), since the true $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are unknown, the estimated densities $\hat{p}(\boldsymbol{Y})$ and $\hat{p}'(\boldsymbol{Y})$ are used to calculate the PE divergence. In the relative density-ratio estimation method, instead of estimating two distributions $\hat{p}(\boldsymbol{Y})$ and $\hat{p}'(\boldsymbol{Y})$ respectively (a harder problem), we only estimate one statistic, the density-ratio $g(\boldsymbol{Y};\boldsymbol{\theta}) = \hat{p}(\boldsymbol{Y})\ /\hat{p}'(\boldsymbol{Y})$, through Gaussian kernel model [36]

$$g(\boldsymbol{Y};\boldsymbol{\theta}) = \sum_{l=1}^{n} \theta_l K(\boldsymbol{Y}, \boldsymbol{Y}_l) \tag{4}$$

where $\boldsymbol{\theta}$ is an n dimensional parameter to be learnt from the data samples so that the PE divergence between $p(\boldsymbol{Y})$ and $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ is minimized; and $K(\boldsymbol{Y},\boldsymbol{Y}_l)$ is the Gaussian kernel function evaluated at $\boldsymbol{Y}_l$.

After the density-ratio estimator $\hat{g}(\boldsymbol{Y})$ is computed using the estimated $\widehat{\boldsymbol{\theta}}$, the PE divergence can be constructed as equation (5) [30].

$$\widehat{PE} = -\frac{1}{2n}\sum_{j=1}^{n}\hat{g}\left(\boldsymbol{Y}'_j\right)^2 + \frac{1}{n}\sum_{j=1}^{n}\hat{g}\left(\boldsymbol{Y}'_j\right) - \frac{1}{2} \tag{5}$$

If we consider the α-relative PE-divergence PE_α for $0 \leq \alpha < 1$, the symmetrized PE divergence is given as

$$PE_\alpha(P||P') + PE_\alpha(P'||P) \tag{6}$$

where $PE_\alpha(P||P') = PE(P||\alpha P + (1-\alpha)P')$ and α is called the "smoother" as α gets larger [37].

According to Reference [30] the introduction of a relative density-ratio provides a solution for the unbounded density-ratio for better estimation. The adopted density-ratio estimation method is also known as relative unconstrained least-squares importance fitting (RuLSIF). Compared with other change-point detection methods, RuLSIF has several advantages for PV installation detection. First, RuLSIF is parameter-free. We only need to control the time window length k, as shown in equation (2). Second, RuLSIF estimates one density-ratio instead of two density functions, which is computationally efficient and

substantially easier [30]. Finally, RuLSIF is known for its optimal non-convergence rate and robustness compared with other time-series-based methods [30].

2.3.3 PV Detection Identification

The change-point detection algorithm discussed above can detect abnormalities in a customer's energy consumption history caused by the PV installation. However, other customer behaviors such as introducing a new EV or a sudden drop of temperature will also cause abrupt energy consumption abnormalities and thus be detected and marked by a change point. As a result, once an abnormality is detected, a statistical inference must be constructed to further verify whether the sudden change of customer behavior is caused by the installation of a PV system.

2.3.3.1 Typical Load Profile

The typical load profile, which summarizes the customer's energy consumption pattern, plays a fundamental role in a utility's daily operation. We introduce a daily TLP to compare a customer's power consumption patterns before and after the change point. Let us assume a smart meter collects f readings per day. The daily TLP of a specific customer can be represented by a vector $\boldsymbol{V}_{TPL} \in \mathbb{R}^f$. Given a time window of n days, the TLP for the customer can be computed using equation (7).

$$\boldsymbol{V}_{TLP}(\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D}(i) \in \mathbb{R}^f \tag{7}$$

Due to the fact that most smart meters are installed at the residential level, the random behaviors of homeowners may cause spikes along their energy consumption history. These

spikes introduce significant noises to TLP estimation in equation (7). In order to filter out unnecessary noises, we use the Gaussian kernel density method to estimate the TLP. Kernel density estimation is a non-parametric algorithm originally used for probability density function estimation. Since kernel density estimators asymptotically converge to any density function with sufficient samples, it is a very general estimation method [38] and is robust for a variety of TLP shapes. Compared with simply taking the mean value in (7), Gaussian kernel approach returns a much smoother TLP with less noise and requires less space to store. In our study, the TLP curve is treated as a probability density function and Gaussian kernels are used to estimate the TLP. The estimated density function $\hat{f}(x)$ with m kernels can be computed by equation (8):

$$\hat{f}(x) = \sum_{i=1}^{m} w_i K(x - x_i) \tag{8}$$

where $K(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-x_i)^2}{2\sigma^2})$ is the Gaussian probability density function with mean x_i and variance σ^2, w_i is the weight of each Gaussian kernel that satisfies $\sum w_i = 1$.

Figure 5 shows two TLPs of a customer before and after a PV system was installed. The blue curves are TLPs computed using equation (7). The red curves correspond to TLPs smoothed by the Gaussian density estimation method. It is clear that the Gaussian-shaped kernel serves to smooth out the noises in TLP.

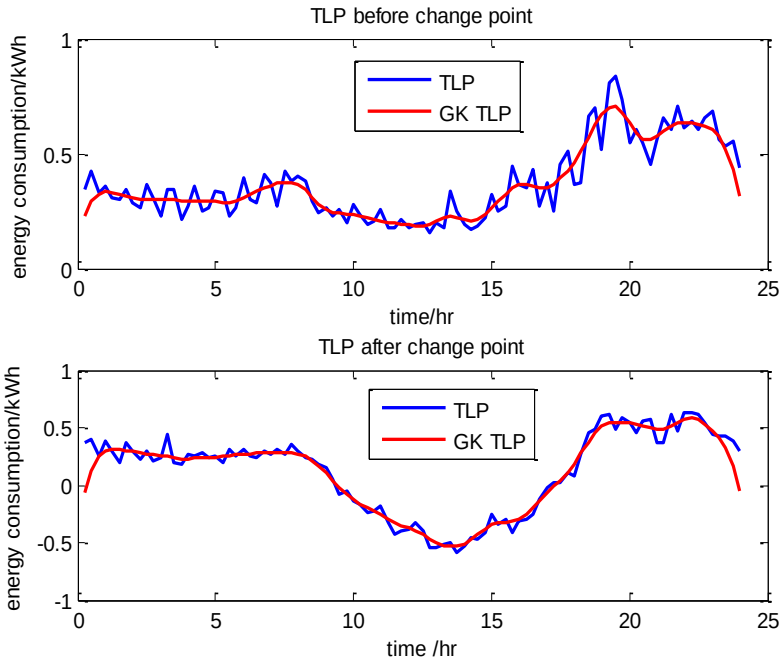

Figure 5. Gaussian kernel-based TLP.

2.3.3.2 Statistical Hypothesis Test with Spearman's Rank

When an abnormal customer behavior is detected, it is crucial for utilities to verify whether the abnormality is caused by a PV installation. Instead of issuing a field work order and on-site inspection, we construct a TLP-based hypothesis test to verify the existence of an unauthorized PV system. Specifically, we construct the null hypothesis (H_0) as the following statement: "There is no unauthorized PV system installed by the customer." In other words, we generally assume that there is no unauthorized PV installation unless evidence strongly indicates otherwise.

Similar to a customer's TLP, $\boldsymbol{V}_{TPL} \in \mathbb{R}^f$, we define $\boldsymbol{V}_{PV} \in \mathbb{R}^f$ as a standard TLP of a local PV system. $\boldsymbol{V}_{PV}$ records the standard daily energy output of the local PV systems with rated power equal to 1 kW. Let $\Delta\boldsymbol{V}_{TLP} \in \mathbb{R}^f$ denote the difference of TLPs before and after the change point. If the detected change point is caused by an unauthorized PV system,

we have $\Delta \boldsymbol{V}_{TLP} = p\boldsymbol{V}_{PV}$, where p is the size of the unauthorized PV system. Otherwise, we will be unable to find a constant p so that $\Delta \boldsymbol{V}_{TLP} = p\boldsymbol{V}_{PV}$ is true.

Let us define

$$\Delta \boldsymbol{V}_{TLP} = \boldsymbol{X} = (x_1, x_2, x_3, \dots, x_f) \tag{9}$$

$$\boldsymbol{V}_{PV} = \boldsymbol{Y} = (y_1, y_2, y_3, \dots, y_f) \tag{10}$$

Then, the original hypothesis test can be rephrased as:

$$\begin{cases} H_0: \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are not positively correlated} \\ H_1: \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are positively correlated} \end{cases} \tag{11}$$

2.3.3.3 Spearman's Rank and Permutation Test

Pearson product-moment correlation (Pearson's r) and Spearman's rank correlation coefficient (Spearman's rank) are the most commonly used metrics to quantify the correlation between two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ [39]. However, the difference between the two methods lies in that Pearson's r assumes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are normally distributed, while Spearman's rank does not have any requirement on the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. We adopt Spearman's rank (r_s) because the distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in (9) and (10) are not normal. The Spearman's rank coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be computed using equation (12) [40].

$$r_s = 1 - \frac{6(\sum d_i^2)}{n(n^2 - 1)} \tag{12}$$

where r_s is the Spearman's rank coefficient ($-1 \leq r_s \leq 1$). When $|r_s|$ is close to 1, it indicates a strong linear relationship between the two distributions, and 0 otherwise. n is the number of (x_i, y_i) pairs in observation which, in our case, $n = f$ and $d_i = x_i - y_i$. Since r_s quantifies the strength of the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, an interpreting table developed by Hinkle [40] is usually used for interpreting the physical meaning of r_s [41].

Table 1 – Rule of Thumb for Interpreting the Size of A Correlation Coefficient.

Size of Correlation	Interpretation
90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	Negligible correlation

In our hypothesis test, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not normally distributed, we cannot use a t-test to acquire an accurate p-value through the student distribution. Instead of using t-test, we adopt the permutation test. Permutation test (a.k.a. randomization test) is a very general approach to test a statistical hypothesis, where the distribution of the observations under the null hypothesis need not be known to obtain the p-value [42].

The existence of an unauthorized PV system will drive r_s close to 1. Hence, we can further rephrase the original null hypothesis in (11) as $H_0: r_s = 0$. Next, we select a significance level α and compute the p-value through the permutation test. For f pairs of (x_i, y_i) listed in (9-10), the total number of permutation sets is 2^f. Let $r_{s,i}$ stand for the Spearman's rank coefficient for permutation set π_i and $r_{s,0}$ for the observed Spearman's rank coefficient of permutation π_i. Since we want to test whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are positively

correlated, the permutation test is no longer a two-tailed test but an upper-tailed test. Therefore, the corresponding test procedure can be decomposed using the following three steps:

Step 1 Generate all possible permutation sets [43]: $\pi_1, \pi_2, \dots, \pi_{2^f}$.

Step 2 Compute the Spearman's rank for all sets: $r_{s,1}\ r_{s,2}, \dots, r_{s,2^f}$.

Step 3 Construct an empirical cumulative distribution [42]:

$$\hat{p}\left(r_s \le r_{s,0}\right) = \frac{1}{2^n}\sum_{i=1}^{2^n} 1(r_{s,i} \le r_{s,0}) \tag{13}$$

where $\hat{p}$ is the cumulative density function of the estimated Spearman's rank coefficient. $1(s)$ is an indicator function which takes value 1 if statement s is true and 0 otherwise. In practice, when the number of (x_i, y_i) pairs is generally large (in our case $f = 96$), it is difficult to generate all possible 2^f permutations. As a result, bootstrap sampling must be implemented. For the significance level of α=0.05, according to Reference [33], 10,000 bootstrap samples are recommended.

Given a preset significance level α, we reject the null hypothesis, if $\hat{p} \le \alpha$. In other words, $\hat{p} \le \alpha$ indicates that there is a very good chance the detected customer has installed an unauthorized PV system.

2.4 Unauthorized PV System Estimation

Among all PV parameters, the size or the rated power of the PV system p is the most important. However, as a parameter estimation problem, a good estimation of p is difficult when only smart meter measurements are available. This is because the PV output is

strongly affected by weather conditions such as local solar irradiance and cloud cover. CCI obtained from satellite images contains information on cloud extent and optical thickness [44]. To be more specific, CCI is defined as an integer ranging from 0 to 8, where 0 stands for clear-sky day and 8 stands for heavily-clouded day.

In this section, we select a residential PV system and record its output for 91 consecutive days, as shown in 0-1. The local CCIs for the corresponding days are shown in 0-2. We can see that on high CCI days, the PV output is generally small, and vice versa. The correlation between the PV daily output and the CCI is -0.8554, which indicates high linear correlation between the two. This meets our expectation that cloudy skies lead to lower PV output.

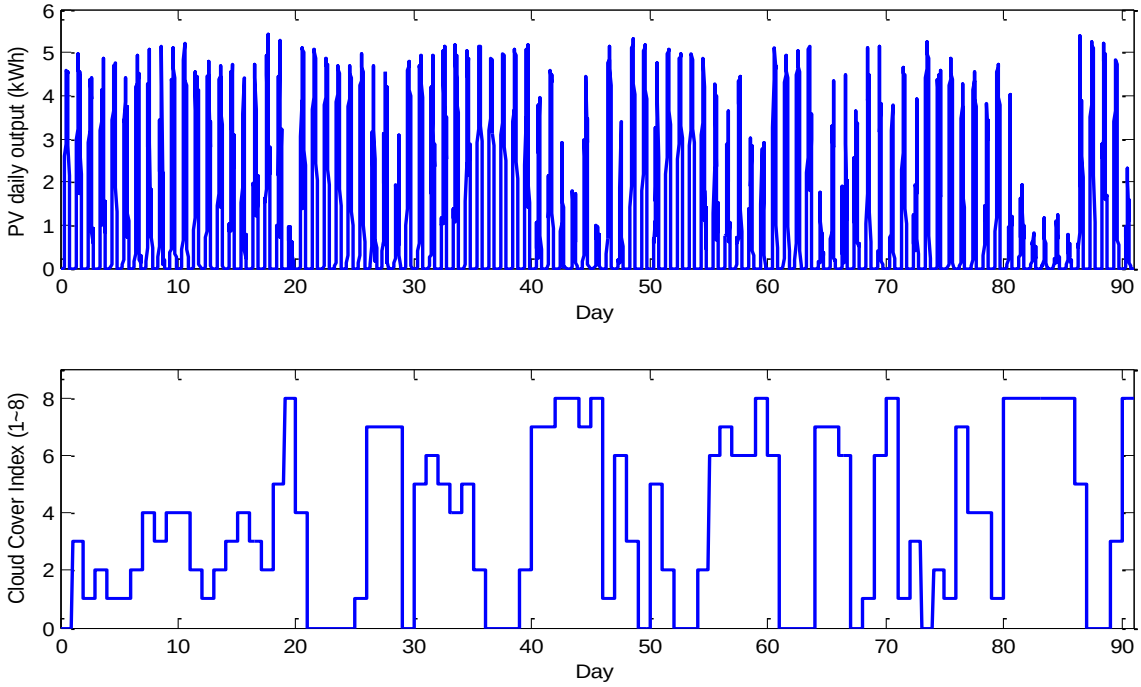

Figure 6. PV output and corresponding CCIs for 91 days.

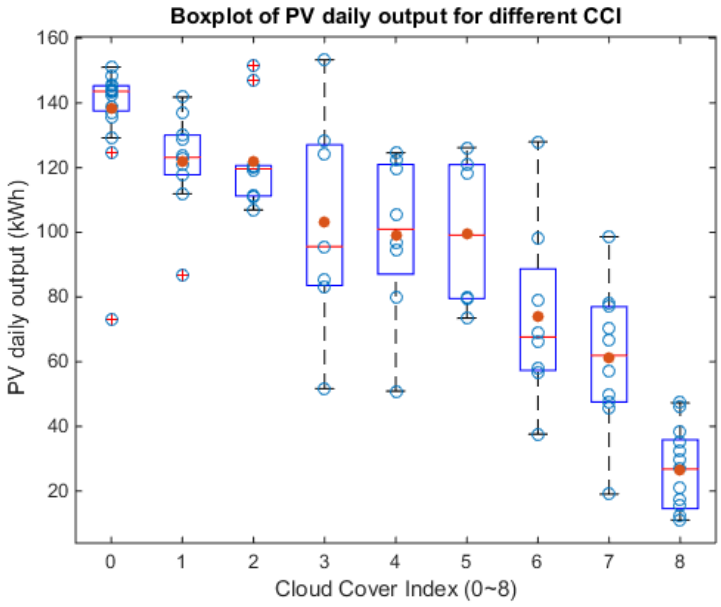

Figure 7. Boxplot of PV daily output vs. local CCI.

In order to visualize the correlation between CCI and PV output, the boxplot (a.k.a. box and whisker diagram) of PV daily output condition on the CCIs is shown in Figure 7 using the previous data. According to the boxplot definition [45], the central red mark is the median, the edges of the box are the first and third quartiles, and the red cross stands for outliers. From Figure 7, we can see that the PV output variance increases as the CCI increases from 0 and decreases when CCI approaches 8. This phenomenon can be explained by the fact that when CCI is in the middle range, the sky is partially covered by clouds, and the passing of clouds above the solar panel may lead to a huge variance on PV output. In order to obtain an accurate PV size estimation, only days with low CCI can be used, where the PV output has a small variance, near its rated output.

Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ stand for the smart meter readings before and after the PV installation respectively. Let $\widetilde{\boldsymbol{D}}_2$ be an adjusted $\boldsymbol{D}_2$ according to local CCIs and radiance. For a specific day k, $\widetilde{\boldsymbol{D}}_2(k)$ can be computed using (14).

$$\widetilde{\boldsymbol{D}}_2(k) = \boldsymbol{D}_2(k) - p \times p_{CCI}(k) \times \boldsymbol{V}_{PV} \tag{14}$$

where p is the size of the PV system, $p_{CCI}(k)$ is the adjustment coefficient related to the local CCI and radiance on day k, which increases as CCI increases. In practice, p_{CCI} can be estimated based on empirical distribution of a PV output condition on the local CCI. Then, the PV size estimation problem becomes choosing the best constant p that minimizes (15).

$$min: \left\| \boldsymbol{V}_{TLP}(\boldsymbol{D}_1) - \boldsymbol{V}_{TLP}(\widetilde{\boldsymbol{D}}_2) \right\|^2 \tag{15}$$

where $\boldsymbol{V}_{TLP}(\boldsymbol{D}_1)$ and $\boldsymbol{V}_{TLP}(\widetilde{\boldsymbol{D}}_2)$ stand for the typical load profiles computed using $\boldsymbol{D}_1$ and $\widetilde{\boldsymbol{D}}_2$.

2.5 Real Case Analysis

In this section, we investigate the performance of our method on real data sets. The data contain a rich source of disaggregated customer energy consumption. In order to show the robustness of the proposed method, a representative subset of the data described in Section II is used which includes three distinct scenarios:

- Scenario 1: Customer A has installed an unauthorized PV system;
- Scenario 2: Customer B has bought a new EV and experienced a major weather change;
- Scenario 3: Customer C has no abnormal behavior.

We expect that our proposed algorithm will only identify the customer in scenario 1, where an unauthorized PV system exists.

2.5.1 Change-point Detection Screening

The proposed change-point detection algorithm will pick up energy abnormality efficiently when historical smart meter data are available. The real case study shows that only customer C in scenario 3, who does not have any abnormal energy consumption behaviors, can pass our change-point detection screening.

Scenario 1: An unauthorized PV system is installed

In scenario 1, a smart meter monitored the aggregated power consumption of customer A for 91 consecutive days with 8736 measurements, as shown in Figure 8-1. The negative values in Figure 8-1 stand for the PV system back feeding to the grid. The unauthorized PV system was installed on the 41th day and the PV output is recorded by a separate meter as shown in Figure 8-2. The reading of this meter is invisible to the local utility.

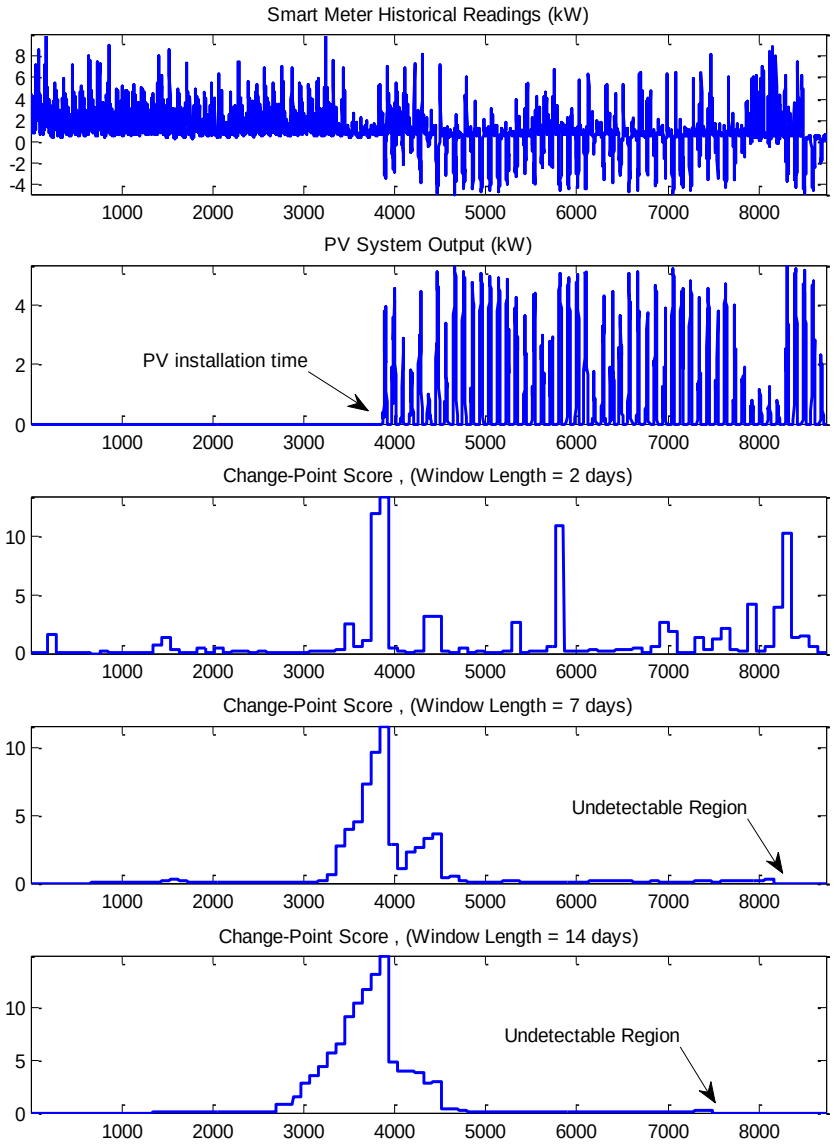

Figure 8. Change-point detection screening for an unauthorized PV installation.

Given the parameter-free nature of RuLSIF, analysts only need to determine the estimation window length k as in equation (2). The performance of the change-point detection algorithm relies on a proper choice of k. The algorithm takes the aggregated data in Figure 8-1 as inputs, and returns the PE divergence scores in Figure 8-3, Figure 8-4 and Figure 8-5, each with a different time window length (2 days, 7 days and 14 days). Due to the smart meter data structure formulated in equation (1) and (2), the algorithm will leave

two blind detection periods located at the beginning and the end of the time series stream as shown in Figure 8-4 and Figure 8-5. The undetectable period length equals to the length of the estimation window k. In other words, the algorithm cannot detect a newly installed PV system until k days after the initial installation. From Figure 8-3, Figure 8-4, and Figure 8-5, we see that a shorter estimation window will enable the detection of some short term changes in customer behavior and also minimize the undetectable period at the expense of lower index stability. However, the installation of a PV system is not likely to be a short-term activity; a longer estimation window can increase the robustness of the algorithm. Compared with Figure 8-4 and Figure 8-5, Figure 8-3 is generated with a much shorter time window and its PE divergence score is less stable. Therefore, a balance must be maintained when choosing a proper time window. In our study, we set an appropriate estimation window length as 7 days.

Scenario 2: A new EV and load fluctuations caused by weather changes

In Scenario 2, no unauthorized PV is presented during the 91-day study period. However, a new EV was introduced at the 51th day and the customer also experienced a sudden temperature change at the 82th day. From Figure 9, the change-point detection algorithm picks up two change-points when the EV was introduced and when the temperature fluctuated.

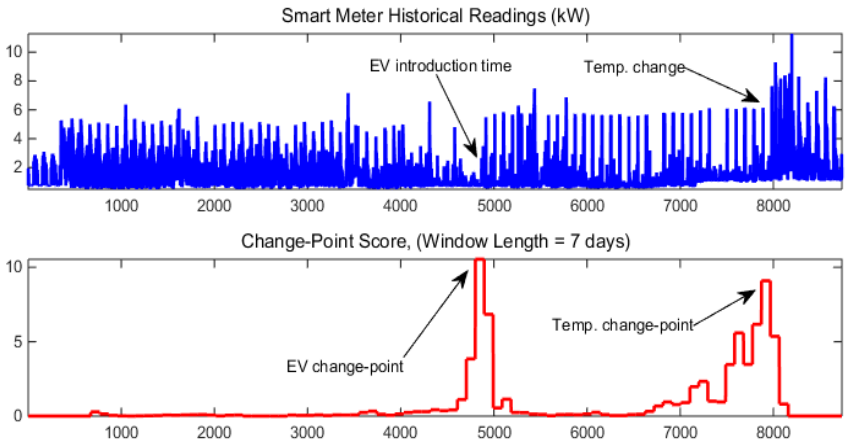

Figure 9. Change-point detection screening for a new EV and temperature changes.

Scenario 3: Customer without abnormal behaviors

In Scenario 3, there is no PV, EV introduction or huge temperature fluctuations, as shown in Figure 10. The change-point detection algorithm does not pick up any significant change point and the PE divergence scores are consistently below 2.5. As a result, the customer in scenario 3 passes our change-point screening test (no abnormality has been detected).

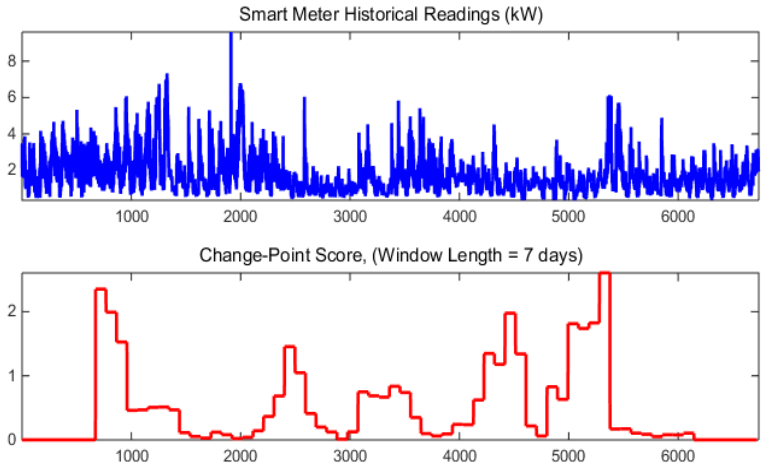

Figure 10. Change-point detection screening for customer without abnormal behaviors.

2.5.2 PV System Verification

In the previous step, only the customer in scenario 3 passes the change-point screening test, which leaves us with customers A and B of scenarios 1 and 2. In the second step, we use the statistical inference constructed in Section IV to identify customers without an unauthorized PV system, but who fail the screening test as in scenario 2. We first create the Gaussian kernel-based TLPs before and after each detected change point and compute their differences. Next, we conduct the permutation test with Spearman's rank coefficient to verify the existence of an unauthorized PV system.

In this study, the standard local PV system output profile $\boldsymbol{V}_{PV}$ is approximately estimated by taking the normalized output of 40 local PV systems on a cloud-free day, as shown in Figure 11-1. The $\Delta\boldsymbol{V}_{TLP}$ for three change points in scenario 1 and scenario 2, as shown in Figure 11-2, Figure 11-3 and Figure 11-4, are computed using equation (7) and (8). In this study, we choose 10 days as the time window to create the TLPs. All the TLPs in Figure 11 are normalized and smoothed by Gaussian kernel method.

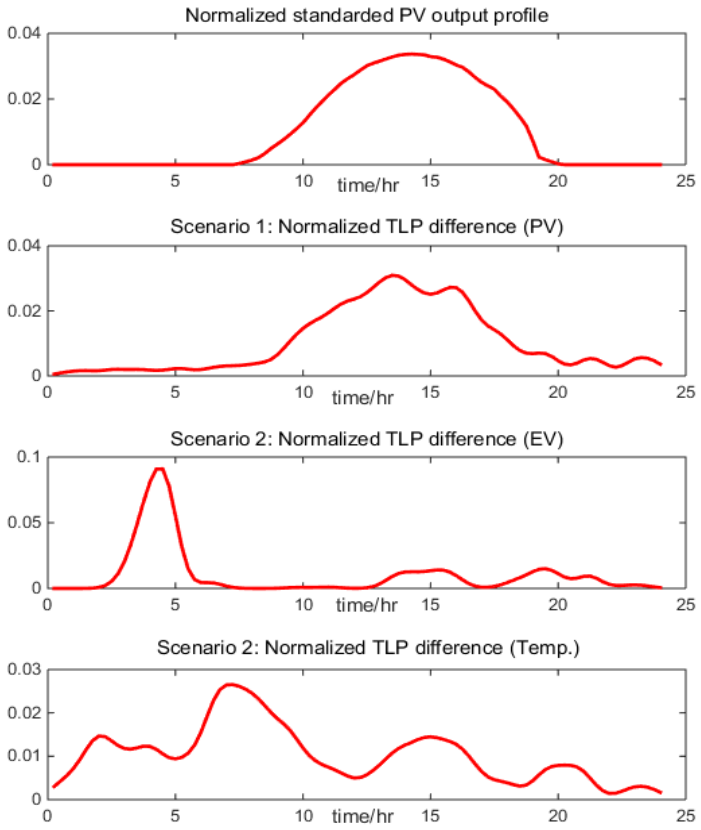

Figure 11. Gaussian kernel-based typical load profiles.

Based on Figure 11, we perform correlation strength analysis between the standard PV output $\boldsymbol{V}_{PV}$ (Figure 11-1) and $\Delta\boldsymbol{V}_{TLP}$ of each detected change point (Figure 11-2, Figure 11-3, and Figure 11-4). Table 2 lists the Pearson's r and the Spearman's rank coefficient for each change point. In 0 only the customer in scenario 1 returns high Pearson's r and Spearman's rank coefficient which strongly indicate the existence of an unauthorized PV system. Moreover, with a choice of significance level $\alpha = 0.05$, we only reject the null hypothesis in scenario 1 where the p-value is much less than α. In scenario 2, both the p-values for the EV case and temperature case are much greater than α, which means we cannot reject our null hypothesis: there is no unauthorized PV system installation

for customer B in scenario 2. After the verification step, we only accept the alternative hypothesis in scenario 1, where an unauthorized PV system truly exists.

Table 2 – Correlation Strength Analysis.

Change Point	Pearson's r	Spearman's rank coefficient	
	r	r_s	p-value
Scenario 1 (PV)	0.9205	0.8351	3.9414e-26
Scenario 2 (EV)	0.1290	-0.0315	0.7609
Scenario 2 (temp.)	0.0817	-0.0754	0.4651

2.5.3 Algorithm Sensitivity Analysis

In order to test the robustness of the proposed algorithm, we perform a sensitivity study for both the change-point detection algorithm and the statistical inference against the PV system size. To achieve this, we need to block all other factors which may influence our result except the PV system size. As a result, we pick the same customer with the fixed energy consumption but manually scale the output of the PV system from 100% to 10% of its original output. Let C be the energy consumption of a house and S be the PV output from the home solar system. $V = C - rS$ is the energy measurement visible to us, where r is the PV size scaling factor ranging from 100% to 10%, as shown in Table 3. The goodness of the change-point detection in Table 3 is a measurement used to quantify how confident we are about the detection [30]. The smaller the goodness value, the more reliable the detection result. No change point is detected if the goodness of the detection is above one. If we consider a significance level $\alpha = 0.05$, both the detection algorithm and the statistical inference show great sensitivity. Both of them fail only in the case where we maintain the

energy consumption of the customer and scale down the PV system to 10% of its original size.

Table 3 – Sensitivity Analysis.

Scaling Factor r	Change-Point Detection		Permutation Test $\alpha = 0.05$	
	Detection	Goodness	r_s	p-value
100%	yes	0.3114	0.8351	3.9414e-26
90%	yes	0.3186	0.8268	3.1698e-25
80%	yes	0.3334	0.8169	3.4432e-24
70%	yes	0.3540	0.7985	1.9597e-22
60%	yes	0.3833	0.7686	6.1399e-20
50%	yes	0.4135	0.7197	1.4309e-16
40%	yes	0.4593	0.6297	6.2748e-12
30%	yes	0.5568	0.4779	8.4663e-07
20%	yes	0.8000	0.2481	0.0148
10%	no	1.5076	-0.0358	0.7291

2.5.4 PV Size Estimation

The third step is estimating the unauthorized PV system size. In Section V, we show that the PV output is strongly correlated with the local CCI. For simplicity, we only choose the PV output data when the local CCI is zero (clear sky days) using equation (8) and (9). For a 5kW PV system, we get the estimated PV size of 4.7912kW using the CCI information (p_{CCI} is set at 1.07 according to the empirical PV output distribution condition on local CCI and irradiance). Without CCI information, data with high CCI are also used, which leads to a PV size estimation of 2.7771kW. In fact, due to the strong correlation between PV output and CCI, it is almost impossible to get an accurate PV size estimate without the local CCI.

2.6 Conclusion

In this chapter, we propose a data-driven approach for residential PV detection, verification, and estimation. The proposed method consists of three steps. On the first step, the unauthorized PV installation events and other abnormal customer behaviors are detected through change-point detection. On the second step, permutation tests based on Spearman's rank coefficient are constructed to verify the existence of unauthorized PV systems. On the last step, the PV system size is estimated with the help of the local weather information. A realistic data study demonstrates the effectiveness and robustness of the proposed method. In the future, we would like to expand our detection and estimation to other critical load components, such as EV and temperature-related loads. The disaggregation and detection of these critical load components plays an important role for utilities to ensure safe and reliable operations.

CHAPTER 3. ELECTRICAL VEHICLE MODELING

As the electric vehicle becomes a significant component of electricity loads, an accurate and valid model for EV charging demand is the key to enabling accurate load forecasting, demand response, system planning, and several other important applications. We propose a data-driven queuing model for residential EV charging demand by performing big data analytics on smart meter measurements. The data-driven model captures the non-homogeneity and periodicity of the residential EV charging behavior through a self-service queue with a periodic and non-homogeneous Poisson arrival rate, an empirical distribution for charging duration and a finite calling population. Upon parameter estimation, we further validate the model by comparing the simulated data series with real measurements. The hypothesis test shows the proposed model accurately captures the charging behavior. We further acquire the long-run average steady state probabilities and simultaneous rate of the EV charging demand through simulation output analysis.

3.1 State-of-the-art Models for Electrical Vehicle Charging Demand

Electric vehicles (EVs) draw and store energy from an electric grid to supply propulsive energy for the vehicle [46]. Since the US federal government highlighted electricity as a promising alternative to petroleum in the transportation sector in 2009 [47], the strong policy support has made US the leader of EV market. As of September 2014, the United States has the largest fleet of highway-capable EVs in the world, with about 260,000 plug-in electric cars sold since 2008 [48].

Many researchers have shown that in a high EV penetration environment, uncoordinated EV charging behavior could have a significant impact on distribution grids, especially at the residential level [49]-[50]. Meanwhile, with a proper control strategy, the battery of the EV could potentially provide additional services to the grid through demand controls, such as flattening the peak load, providing voltage support and frequency regulation. In order to achieve these goals, it is crucial to develop an advanced model that captures the charging behavior of EVs for both operational and planning purposes.

Various research papers [51]-[55] model the EV charging process as a queuing system. In reference [51], the EV charging time and duration are determined in a deterministic manner by some market signals and a fixed distance distribution. In reference [52], a $M/M/N_{\max}$ queue is introduced, where the EV arrives as a Poisson process with an exponentially distributed charging time, and N_{max} is the total charging capacity. Reference [53] employs an $M/M/\infty$ queue to capture the fact that residential EV charging is a self-service system. Both reference [52] and [53] assume that the EV charging arrival rate is not related to the number of EVs that are already in charging mode. The $M/M/s$ models in reference [54] and [55] are based on the assumption that the arrival process of an EV charging event is a homogeneous Poisson process with a constant rate, and that the charging duration is exponentially distributed. Although we can derive the long-run average properties of the abovementioned models analytically, most of these models are based on some unrealistic assumptions without validation.

Thanks to the widely installed smart meters and corresponding infrastructure, for the first time, researchers and utilities have been able to gain access to the energy consumption patterns of consumers with great resolution and at large scale [56]. In this chapter, we

propose a novel data-driven approach to establish a valid model for residential EV charging demand by applying big data analytics on measurements directly collected from EV charging decks. Although EV charging behaviors are related to factors such as location, customer job, or even the price of gasoline, the smart meter reading alone can be a good indicator, which summarizes all these social factors. The proposed model allows us to capture the non-homogeneity and periodicity of the EV charging demand. Moreover, we estimate the EV charging duration with an empirical pdf generated from the real smart meter data.

The proposed new model does not require any of the pre-assumption mentioned above. In addition, the model can be further utilized by electric utilities for enhanced projection of EV demand and deployment of advanced coordination applications as part of demand response and grid services procurement.

3.2 Stochastic Models of Electrical Vehicle Charging Demand

3.2.1 Data Observation

The key advantage of the data-driven EV model is that the model is supported by real smart meter measurements. The smart meter data not only provide us with the knowledge of residential EV charging patterns, but also plays a vital role in model validation.

Figure 12 depicts some general observations of 37 independent EVs behaviors collected by Pecan Street Inc. [57], Austin, Texas. The data were collected every 15 minutes directly from EV charging decks. In Figure 12-1, Black bars represent charging

behaviors for the 37 EVs; Figure 12-2 shows the number of charging EVs through time; Figure 12-3 visualizes the number of EVs that start charging during each 15 minute time interval; Figure 12-4 shows the energy consumption of all EVs. By observing the four plots, we claim the key of modeling EV charging demand (shown in Figure 12-4) is the modeling of Figure 12-2 through time, which can be further derived from EV charging duration (shown in Figure 12-1) and EV charging arrival rate (shown in Figure 12-3).

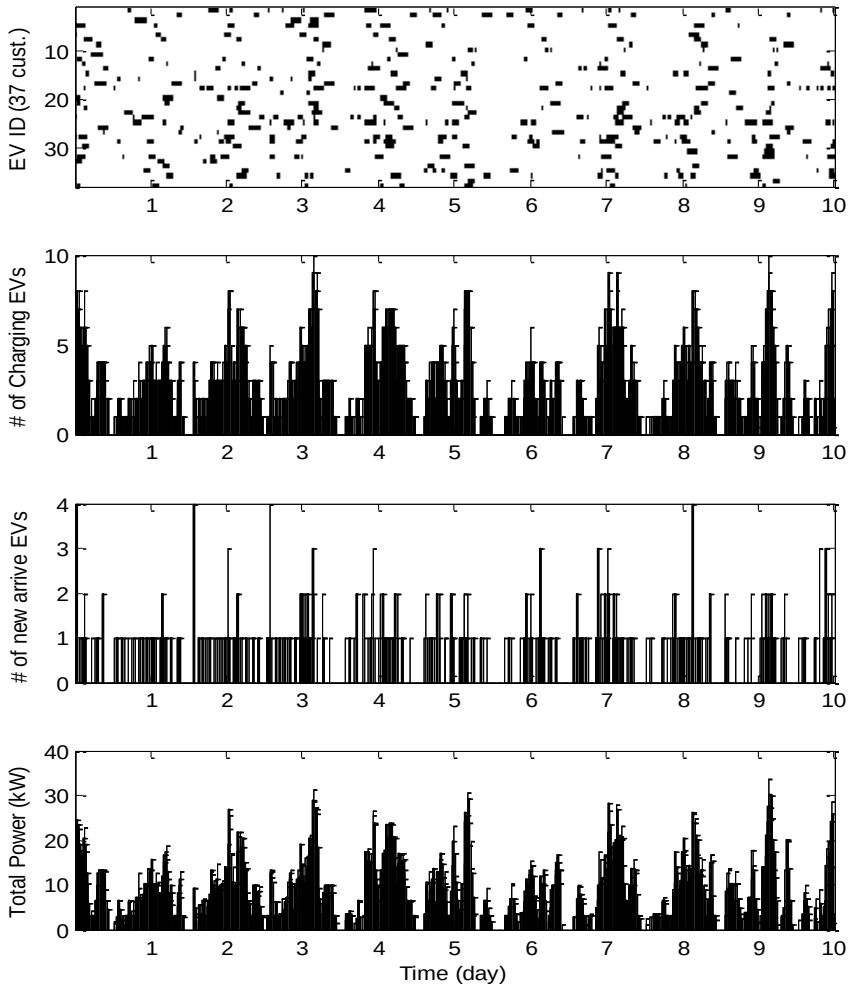

Figure 12. Observation of the EV charging behavior.

3.2.2 A General $M_1/M_2/\infty/N_{max}$ Model

The $M_1/M_2/\infty/N_{max}$ queue is the most widely adopted stochastic model for EV charging demand. In the model:

- M_1 means that the arrival of EV charging events follow a Poisson process with rate λ;
- M_2 means that the EV charging durations are independently and identically distributed (i.i.d.) with an exponential distribution of rate μ;
- ∞ refers to the infinite number of servers in the queuing system. In other words, the residential EV charging system is a self-service system with no waiting time;
- N_{max} refers to the total number of EVs in the community.

Let $X(t)$ be the number of charging EVs at time t, and the state space of $X(t)$ be S, where $S = \{1, 2, \dots, N_{max}\}$. Then, Figure 13 illustrates the transition diagram of the $M_1/M_2/\infty/N_{max}$ queuing system.

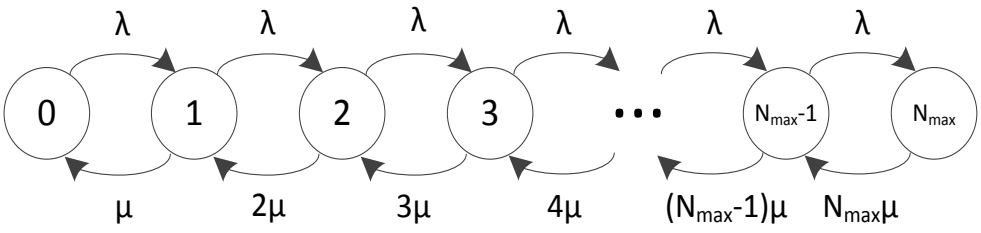

Figure 13. Transition diagram of $M_1/M_2/\infty/N_{max}$ queue.

The advantage of using $M_1/M_2/\infty/N_{max}$ model lies in that researchers can derive the long-run average steady state probabilities of the system analytically. Let P_n denote the

system's long-run average steady state probability of having n EVs charging simultaneously, then P_n can be given directly as

$$P_n = \frac{C_n}{e^{\lambda/\mu}} \tag{16}$$

where $C_n = \frac{\lambda^n}{n!\mu^n}$ and $n = 1, 2, \dots, N_{max}$.

However, some pre-assumptions made by the $M_1/M_2/\infty/N_{max}$ model are not necessarily realistic, which requires further discussions.

3.2.3 $M_1/M_2/\infty/N_{max}$ Queue with Finite Calling Population

To begin with, the $M_1/M_2/\infty/N_{max}$ model assumes the arrival rate of new EV charging events remains the same no matter how many EVs are already in the charging state. However, this is not true as long as the number of EVs is finite. In a community with a finite number of EVs, the potential new arrival rate of new EV charging events decreases as the number of charging EVs increases. In other words, let λ_i be the arrival rate when there are i EVs in the system, for any two integers $\{a, b: 0 \leq \mathrm{a} < \mathrm{b} \leq N_{max}\}$, we have $\lambda_a > \lambda_b$.

To model the finite number of residential EVs, we introduce the finite calling population model [58] for the $M_1/M_2/\infty/N_{max}$ queue. Assume each EV arrives independently according to a Poisson process with rate λ, then $\lambda_i = (N_{max} - i)\lambda$. Figure 14 shows the transition diagram of the system with finite calling population.

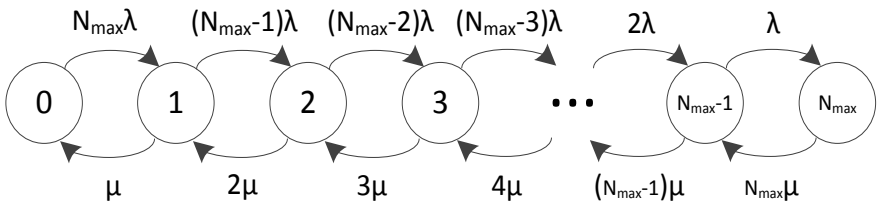

Figure 14. Transition diagram of the finite calling population model.

Another advantage of adopting the finite calling population strategy is making the model scalable and more robust. Under the finite calling population strategy, instead of estimating the behavior of all N_{max} EVs, we estimate the behavior of each EV. As long as the assumption that all EVs behavior independently holds, we could easily fit the model into systems with an arbitrary number of EVs.

3.2.4 Non-homogeneous Poisson Arrival Rate

Another assumption made in $M_1/M_2/\infty/N_{max}$ model is that the arrival rate of EV charging events is a constant throughout time. However, according to Figure 12-3, the arrival rate of EV charging events is not constant through time and has a period of 24 hours. Figure 15 shows the daily average EV charging arrival rate of the 37 residential EVs.

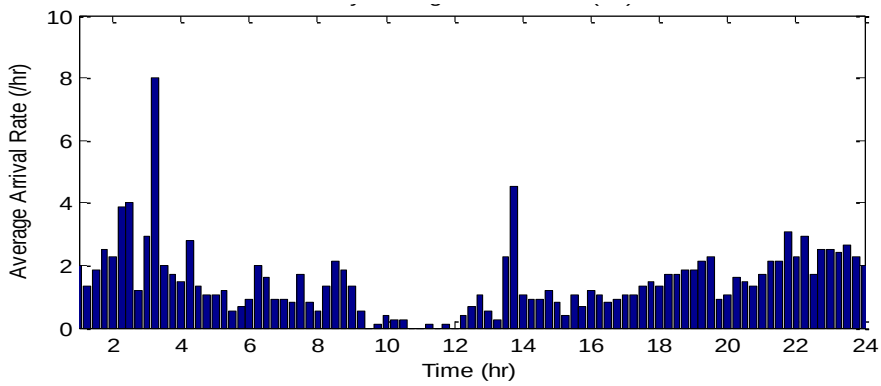

Figure 15. Average daily arrival rate of nonhomogeneous Poisson model.

To illustrate the periodicity of the arrival rate, Figure 16 shows the autocorrelation of the arrival rate with the lag resolution of every 15 minutes. Since the autocorrelation sequence has the same cyclic characteristics as the original arrival rate sequence, Figure 5 can serve to determine and verify the daily periodicity. As expected, the autocorrelation peaks in Figure 16 verify the daily periodicity of the arrival rate.

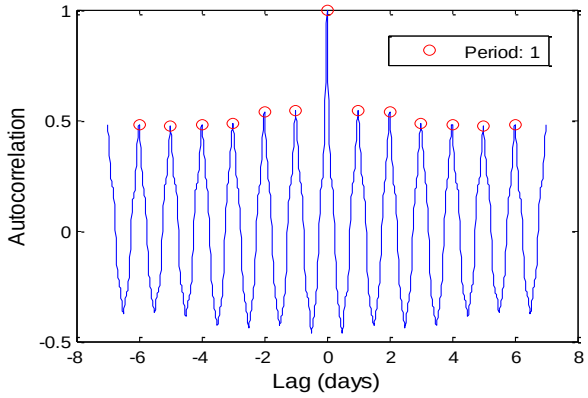

Figure 16. Lag autocorrelation plot of the arrival rate series (30 days).

To capture the time-variant property of EVs, we adopt a non-homogeneous Poisson process with a time dependent rate $\lambda(t)$. Let $m(t) = \int_0^t \lambda(t)dt$, according to the property of non-homogeneous Poisson process, the number of new arrivals from $t = t_0$ to $t = t_1$ follows the Poisson distribution of rate $\lambda = m(t_1) - m(t_0)$.

3.2.5 A General $M_1/G/\infty/N_{max}$ Model

Another assumption made by the $M_1/M_2/\infty/N_{max}$ model is that the charging duration of EVs is exponentially distributed. We will show this assumption is not valid through the memoryless property of the exponential distribution [59].

Assume an EV starts charging at time $t = 0$. Let $P(t > T)$ stand for the probability that the charging duration t is greater than T hours, and $P(t > T + S|t > S)$ the conditional probability of the charging more than $T + S$ hours given S hours of charging. According to the memoryless property of the exponential distribution, $P(t > T) = P(t > T + S|t > S)$. This contradicts the common knowledge of EV charging behavior, since the battery capacity of EVs is limited.

To better model the EV charging duration, we adopt an empirical charging time distribution estimated from real EV charging measurements.

3.3 Model Estimation and Validation

As mentioned in the previous section, the data-driven model developed in this chapter is based on the historical data of 37 residential EVs for two months. One month of data are used for model training and parameter estimation (training data set), and the other month of data model validation (validation data set).

3.3.1 Model Parameter Estimation

We seek to model the residential EV charging behavior through a $M_t/G/\infty/N_{max}$ queue with a finite calling population, where M_t stands for the periodic non-homogeneous arrival rate; G stands for the empirical distribution of EV charging duration; ∞ means the charging system is a self-serve system with no waiting time; and N_{max} is the number of EVs in the community, which is known.

3.3.1.1 Estimation of the non-homogeneous arrival rate

Given the smart meter data resolution, we divide 24 hours of a day into 96 equal time intervals each with the length of Δt, then we treat the non-homogeneous arrival rate as piecewise constant in each time interval.

Let $\lambda(k)$ be the arrival rate of each EV during time interval $((k-1)\Delta t, k\Delta t)$, where k is a discrete integer from 1 to 96. Let $W(k)$ and $N(k)$ be the number of existing and new arrivals of EVs during the time interval. Then $\lambda(k)$ can be estimated through

$$\hat{\lambda}(k) = \frac{N(k)/\Delta t}{N_{max} - W(k-1)}. \tag{17}$$

Figure 17 visualizes the daily average arrival rate for each EV through time using one month of training data.

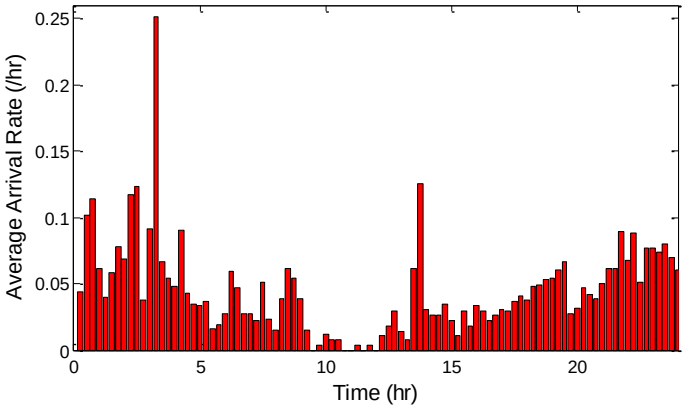

Figure 17. Estimated charging arrival rate per EV.

3.3.1.2 Estimation of EV charging duration

Instead of using exponential distribution, we capture the EV charging duration through an empirical distribution observed from the training data set. Figure 18 shows the empirical probability density function (pdf) of the EV charging duration.

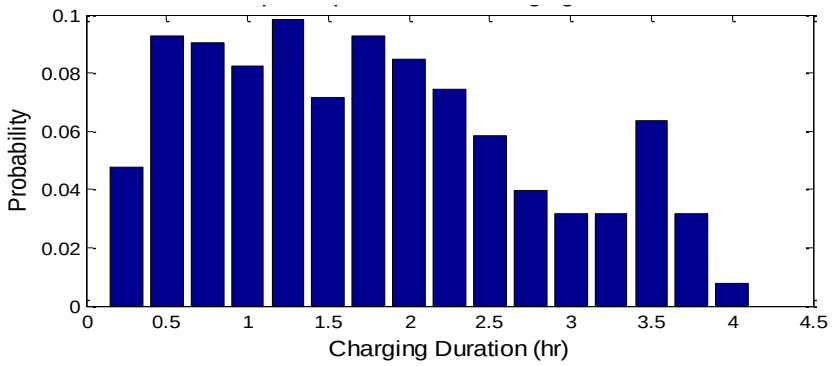

Figure 18. The empirical pdf of the EV charging duration.

3.3.2 Model Validation

Upon the establishment of the model, we further validate it by comparing the simulated data with the validation data. Figure 19 compares our model with real measurements and two other widely used queuing models. From Figure 19-3, we see that if we model the arrival rate as a constant through time, we lose the periodicity and the time variant property of the real measurements. From Figure 19-4, we can see that adopting an exponentially distributed charging duration will distort the true charging behaviors by having charging durations longer than 4 hours, which is unlikely to happen [60]. From Figure 19-2, the simulated data series generated by our model is stable and behavior very similar to the real measurements in Figure 19-1. To validate the model analytically, we run the simulation 100 times (100 replications) each with the length of 100 days. In each replication, the first 10 days' data are trimmed to ensure the data stability.

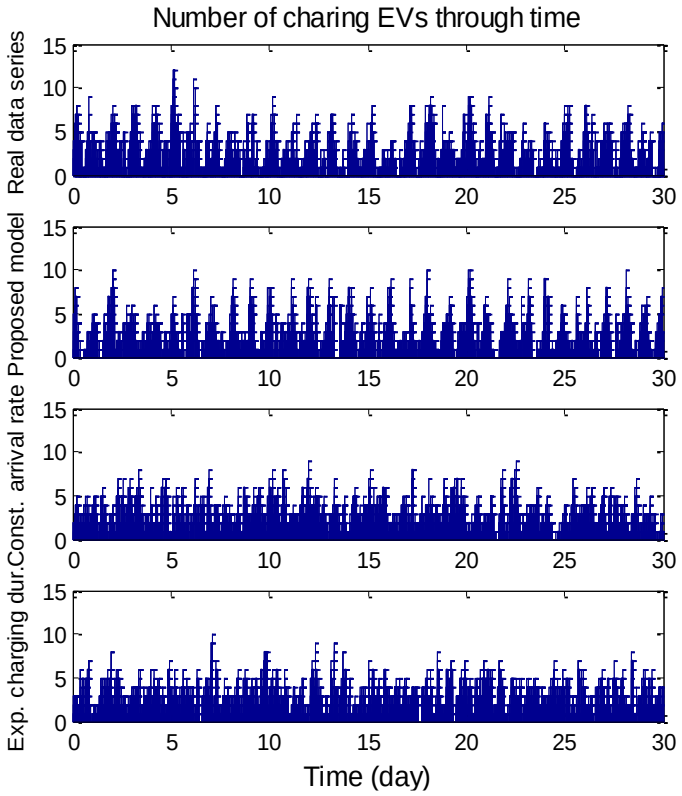

Figure 19. Comparison between simulated and validation data series.

Let $\overline{D}_k$ be the average number of charging EVs during the kth time interval estimated using the validation data, where $k = 1, 2, \dots, 96$. Similarly, let $\widehat{D}_{k,i}$ be the average number of charging EVs during the same time interval estimated by the ith replication. To this end, for each replication, define the difference $G_i = \widehat{D}_{k,i} - \overline{D}_k$, where $i = 1, 2, \dots, 100$.

If the proposed model captures the true EV charging behavior well, G_i should be approximately normally distributed with mean $\mu_g = 0$ and variance σ_g^2 [61]. As a result, we construct a hypothesis test where,

$$\begin{cases} H_0: \mu_g = 0 \\ H_1: \mu_g \neq 0 \end{cases} \tag{18}$$

Under the null hypothesis, the statistic

$$t_{N_2-1} = \frac{\bar{G} - \mu_g}{S_g/\sqrt{N_2}} \tag{19}$$

follows the t distribution with $N_2 - 1$ degrees, where N_2 is the number of the replications, $\bar{G}$ and S_g are sample mean and sample variance [61]. Table 4 compares the statistics $\bar{G}$ and S_g corresponding to the three above mentioned models. It is clear that the proposed model has smaller mean and variance, which means it's a better model of the real EV charging behaviors.

Table 4 – Model Comparison.

Model Type	sample mean $\bar{G}$	sample variance S_g
Constant Arrival Rate Model	-0.0186	1.6199
Constant Charging Rate Mode	-0.0558	0.6345
Proposed Model	-0.0064	0.5109

Given the significance level of $\alpha = 0.05$, we compute the confidence interval for μ_g, which is $(-0.1455, 0.1991)$. Since the interval contains zero, we cannot reject H_0 at the given significance level, which validates the proposed model as a good representation of the EV charging behavior.

3.4 Test Results Analysis

In order to obtain the long-run average steady-state property of the proposed EV charging model, we set the simulation replications to 100, and each replication with the

length of 100 days. Similarly we curtail the first 10 days of each replication due to stability requirements.

3.4.1 Long-run Average Number of Charging EVs

Figure 20 shows the long-run average number of charging EVs throughout a day (blue curve). The 25th and 75th percentiles are also drawn respectively (red and green curves). All three curves suggest that the residential EV charging peak occurs during the night and that the span between 25th and 75th percentiles is relatively small compared to the total EV number of 37.

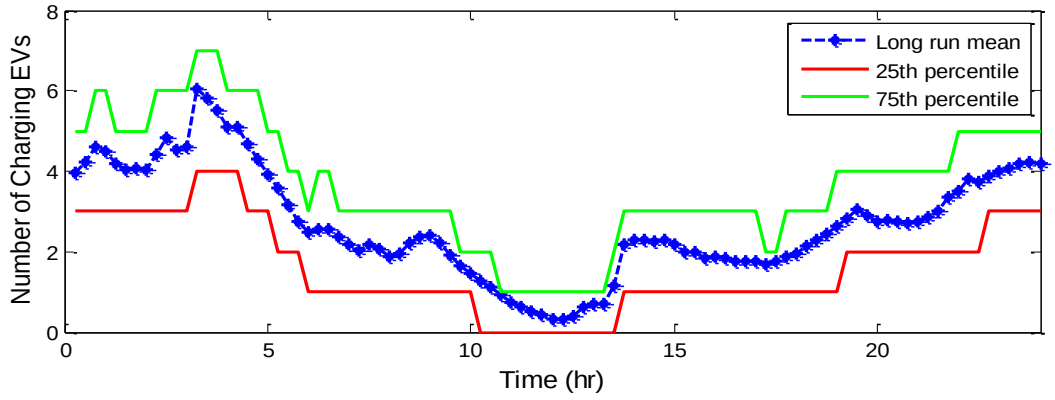

Figure 20. The long-run average, 25th and 75th percentile curves.

3.4.2 Long-run Average Steady State Probabilities

Let a $N_{max} \times 96$ matrix $\boldsymbol{P}$ be the long-run average steady state probability matrix, where $\boldsymbol{P}(n,k)$ denotes the long-run steady state probability of having n EVs charging during time interval k, then for each $k = 1,2,\ldots,96$, we have

$$\sum_{n=1}^{N_{max}} \boldsymbol{P}(n,k) = 1. \tag{20}$$

We visualize the long-run probabilities of the system through Figure 21, where the color in the plot represents the possibility of have n EVs charging at a given time t.

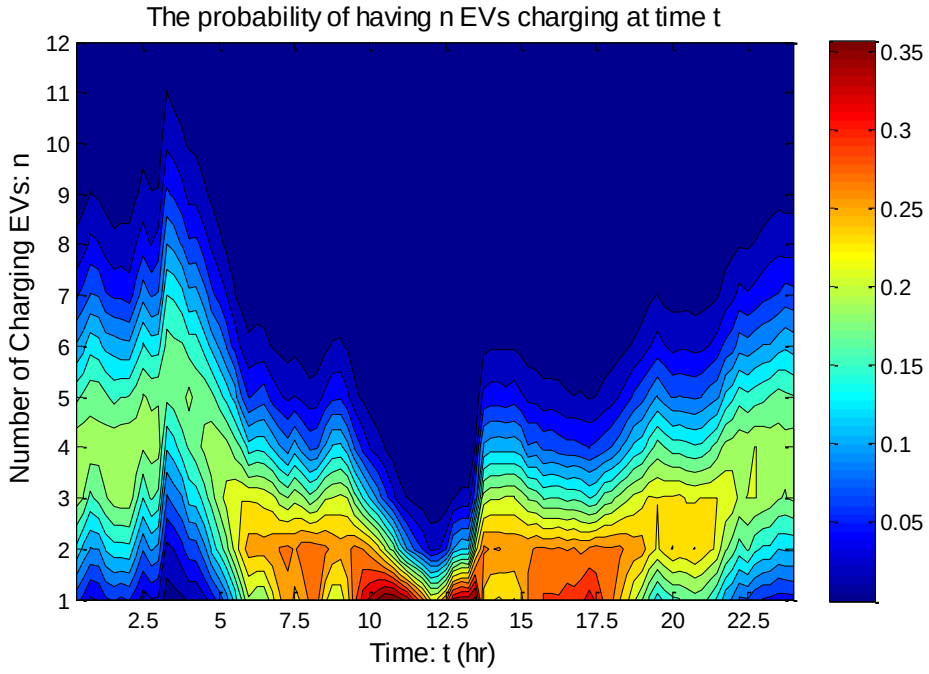

Figure 21. Visualization of the P matrix.

Let ρ be the simultaneous rate of the EV charging load. Define ρ as $\rho = \text{Charging EV number}/\text{Total EV number}$ during the peak EV charging time. Then, the cumulative density function of ρ, which is $P(\rho \leq \rho_0)$, provides essential information to estimate the simultaneous rate of EVs in charging mode. For example, from matrix $\boldsymbol{P}$, we have $P(\rho \leq 12/37) \geq 98.5\%$. This implies that for a community with 37 EVs, even in the worst case, the possibility of having 12 or more EVs charging simultaneously during one day is very slim (less than 1.5%).

3.5 Conclusion

In this chapter, we propose a novel data-driven model for residential EV charging demand. Compared with other queuing models, the proposed model allows us to capture the non-homogeneity and periodicity of EV charging demand, and to estimate the charging duration with an empirical pdf. Upon parameter estimation, we validate the model through hypothesis testing and further acquire the EV charging long-run average probabilities and simultaneous rate through simulation output analysis. The proposed method can be utilized by electric utilities for enhanced projection of EV demand and deployment of advanced coordination applications as part of demand response and grid services procurement.

Further studies may include the analytical deriving of the long-run average steady state statistics for EV charging behavior and the development of corresponding demand response control based on the proposed EV load model.

CHAPTER 4. ADVANCED LOAD MODELING

As part of the ongoing smart grid transformation, smart meters or advanced meter infrastructures have been widely installed, which produce massive amounts of data and information yet unexplored. One of the critical needs for distribution system operation and planning is better modeling of the load. The electric load model in this chapter is described as a mathematical model where the active and reactive power of the load is represented as a function of the voltage. The electrical load model is essential for the operation and planning of the distribution system. In fact, it is also the most commonly used model for the application of voltage conservative reduction, where the utility manually reduces the feeder voltage and temporarily reduces the system load.

We propose a novel time-variant load model through data mining techniques based on a smart meter historical database. Given the data resolution (15 minutes per reading) in the database, the load's P-V and Q-V properties are buried in the spontaneous load changes caused by customer random behaviors. Moreover, massive historical data needs to be labeled, filtered, and clustered before regression. To overcome these barriers, the concept of load condition is introduced. By labeling and clustering smart meter readings, the load's P-V and Q-V properties emerge and enable the establishment of a time-variant load model, which is derived from the traditional ZIP model. The new load modeling method belongs to neither the component-based nor the measurement-based approach, and it is demonstrated using the database collected for the Georgia Tech campus.

4.1 Static Load Model

As one of the essential elements of the smart grid, smart meters have been widely installed in the developed world. It is the first time that utilities and system planners have access to measurements for customers at the building level with great time resolution. The massive historical database created by smart meters contains a wealth of information, which has not been fully explored or exploited. One of the critical needs for enhanced distribution system operations and planning is better modeling of the load. This research proposes a new possibility of building a time-variant load model by implementing data mining techniques on smart meter historical databases.

4.1.1 ZIP Model

From a mathematic point of view, a load model is a formula of the relationship between bus voltage and power (real and reactive) [62]. Compared with the modeling of generators and the transmission system that have been studied in detail, an accurate time-variant load model has been difficult to construct, due to the uncertainty of power system loads and the limitation of data available. Traditionally, the voltage dependency of loads is expressed by exponential or polynomial models with constant coefficients. A time-variant load model is developed based on the traditional ZIP model [63], which is shown in (21) and (22).

$$P = P_0(p_1V^2 + p_2V + p_3) \tag{21}$$

$$Q = Q_0(q_1V^2 + q_2V + q_3) \tag{22}$$

where P and Q stands for the active and reactive power of the load, and $V = \bar{V}/V_0$ is the per unit voltage or the ratio between voltage $\bar{V}$ and its nominal value V_0; P_0 and Q_0 are

active and reactive power of the load at nominal voltage; In ZIP model, p_i and q_i represent the proportions of the corresponding components, which satisfy $\sum p_i = \sum q_i = 1$.

4.1.2 Measurement-based and Component-based Approach

There are two popular approaches to establishing a load model: measurement-based approach [64]-[65] and component-based approach [66]-[67].

The measurement-based approach determines the load model by recording the load responses directly through system voltage-stage tests and actual system transients. Although accurate, the measurement-based approach is costly: testers need to perform specific experiments on real systems by deliberately changing transformer tap positions, which may affect energy quality to customers. Moreover, the measurement-based load modeling method cannot capture the time-variant properties of the load. In other words, the load model built through the measurements only reflects the load's property at the time when those measurements are taken. As a result, it is not realistic to use the daytime load model in midnight load analysis.

The component-based approach estimates the system load's P-V and Q-V properties by aggregating typical load components according to certain ratios, which are also the load's ratios of the typical load components in the system. Instead of taking system measurements, this approach builds a detailed load model in advance for common load components in the studied system, such as televisions in the residential loads or the electric machines in the industrial loads. Hence, the component-based approach avoids costly system tests by taking surveys to determine the ratios of typical load components and building load profiles for each load component. However, the accuracy of this approach

strongly depends on the accuracy of the load components ratios and the specific models built to represent typical load components. As a result, in most cases, the load model built through a component-based approach needs verifications using real system measurements. Table 5 lists the strengths and weaknesses of both the measurement-based method and component-based method.

Table 5 – Comparisons of Measurement-Based and Component-Based Method.

Model Type	Measurement-Based Method	Component-Based Method
Strengths	*Accurate, no need for verification*	*No costly system tests are needed*
Weaknesses	• *Tests are expensive* • *Tests leads to bad power quality* • *Not available on building level* • *Difficult to get 24 hour model*	• *Need pre-modeled models for different load components* • *Large surveys to set those ratios* • *Survey results ≠ load in reality* • *Model needs further validation* • *Difficult to get 24 hour model* • *Does not account for customer behaviors*

4.2 Time-variant Load Model

A static load model does not depend on time [68], and therefore it relates the active and reactive power at a given time to the voltage and /or frequency at the same instant of time. On the other hand, a time-variant load model describes the traditional ZIP load as a function of both voltage and time. Therefore, the time-variant model provides a much more accurate tool for dynamic simulations. In conservative voltage reduction, utility needs to write the system load as a function of voltage and control the feeder voltage to achieve desired peak load alleviating effect. Since the load components of a feeder changes

dramatically through different times of the day and different days of the year, it is essential to have an accurate load model that captures the time variation of the load.

The time-variant model proposed in this chapter consists of multiple static ZIP models, all of which are assigned with a time property. The proposed model has a tree structure that branches through three layers: load type layer, time layer and load condition layer. All the smart meter readings in the database are also labeled correspondingly, as shown in Figure 22.

First Layer: Load Type			
Model Struc.	{ Commercial, Residential, Industrial }		
Data Label	{ Commercial, Residential, Industrial }		
Second Layer: Time			
Model Struc.	{Season}	{Day Type}	{Hour}
Data Label	Spring Summer Fall Winter	Weekday Weekend Holiday	Hr. group 1 Hr. group 2 …
Third Layer: Load Condition			
Model Struc.	{ ZIP Mod. 1, ZIP Mod. 2 … ZIP Mod. K }		
Data Label	{Cond. 1, Cond. 2 … Cond. K }		

Figure 22. Time-Variant Model Structure & Data Label.

On the first layer, all loads are classified into commercial, residential and industrial loads. Ideally, a data mining-based load modeling method does not require a user to specify the load types as long as the load is equipped with smart meters. However, marking the data with load types can help us better understand the different time-variant properties among different load types. This layer also help the utility to incorporate the proposed model into their current business model.

On the second layer, for each individual load, all the smart meter readings are marked with time labels. Later study shows that time labels are good indicators of customer routine behaviors. As a time-variant load model, the proposed model establishes a separate load model for different season, different day type, and even different hour of the day. Please note that the basic time unit for the time-variant load model is an hour. Ideally, we would like to have a distinct load model for every hour of the day under each time label. However, this will lead to a very complicated load model. In fact, it is also unnecessary from a practical point of view. For example, the load model will not likely to change too much from 3am to 4am in early morning. As a result, we merge the hours that share similar load properties, known as an hour group in Figure 22.

On the third layer, smart meter readings with the same time label will further be clustered and marked with different load conditions. These "conditions" might be the result of the control actions of the system controllable elements such as regulators and capacitors, or might be the result of customer behaviors. On this layer, the ZIP model parameters are identified using smart meter measurement data.

4.3 Data-mining-based Load Model

During the load modeling process, data mining and machine learning techniques are implemented. To be specific, Kullback-Leibler (KL) divergence is used to identify and merge different time labels on the second layer; K-subspace method is used to cluster data into different load conditions on the third layer, as shown in Figure 22.

4.3.1 Time Label Identification

Since customers' routine behaviors have a strong correlation with time, the time-variant load model is marked by different seasons of the year, different day type (weekday, weekend, and holiday), and different hour of the day. All data collected by smart meters are marked with the corresponding time labels.

The basic time label unit is set to be one hour. On the one hand, higher resolution of time labels can identify more detailed load behaviors. On the other hand, higher resolution time labels will leave fewer smart meter measurements to each time label for regression. In order to overcome this issue, KL divergence is introduced to group 24 distinct hours of a day into several hour groups, where the load has similar property with each hour group. KL divergences of real power, reactive power, and voltage distributions of all pairs of time labels are evaluated. Then, different time labels with similar routine load behaviors are identified and merged.

KL divergence is a non-symmetric measure of the difference between two distributions. Let $P_1(x)$ and $P_2(x)$ be two distinct distributions, the KL divergence of the two distributions $KL(P_1(x), P_2(x))$ is given by (23) [69].

$$KL(P_1(x), P_2(x)) = \sum_{x \in X} P_1(x) \cdot \log(P_1(x)/P_2(x)) \tag{23}$$

A symmetric variant of KL divergence [70] given by (24) is used in this paper to quantify the divergence of load behaviors throughout different time labels. After computing KL divergence among all pairs of time labels of a day, a KL divergence matrix can be constructed.

$$KL_{sym}\left(P_1(x), P_2(x)\right) = \left[KL\left(P_1(x), P_2(x)\right) + KL\left(P_2(x), P_1(x)\right)\right]/2 \qquad (24)$$

Figure 23 shows the hourly weekday P-V plots for a commercial building on campus for the fall of 2012. KL divergence matrices are computed to merge those hours with highly consistent energy consumption patterns (consistent routine behaviors). Figure 24 visualizes three normalized KL divergence matrices for three distributions respectively: real power, reactive power and voltage. Three specific KL divergence thresholds will be set for the P, Q and V KL divergence matrices to determine which hours can be merged into an hour group. The final hour partition results are the intersection based on the three KL divergence matrices after their individual thresholds have been applied.

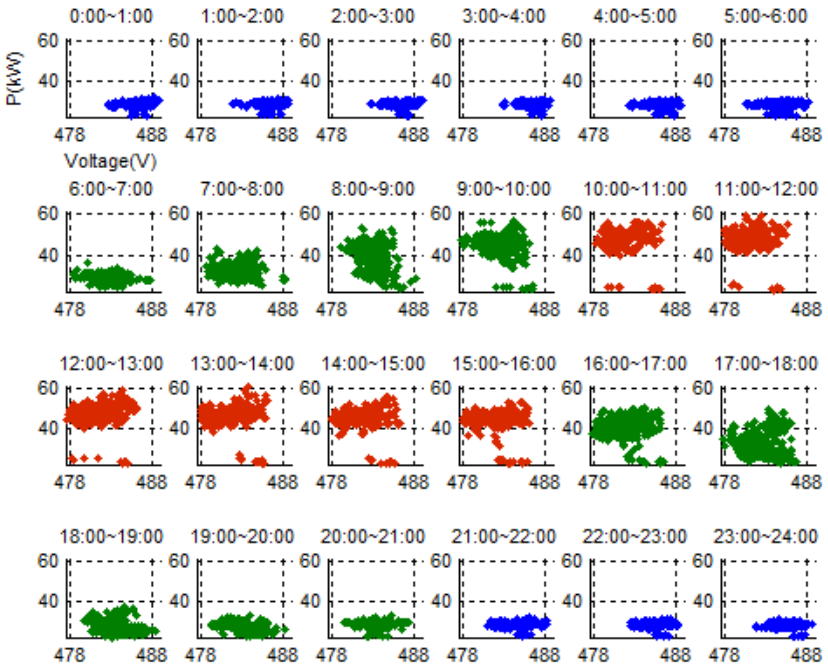

Figure 23. P-V plots for each hour on weekdays for a commercial load.

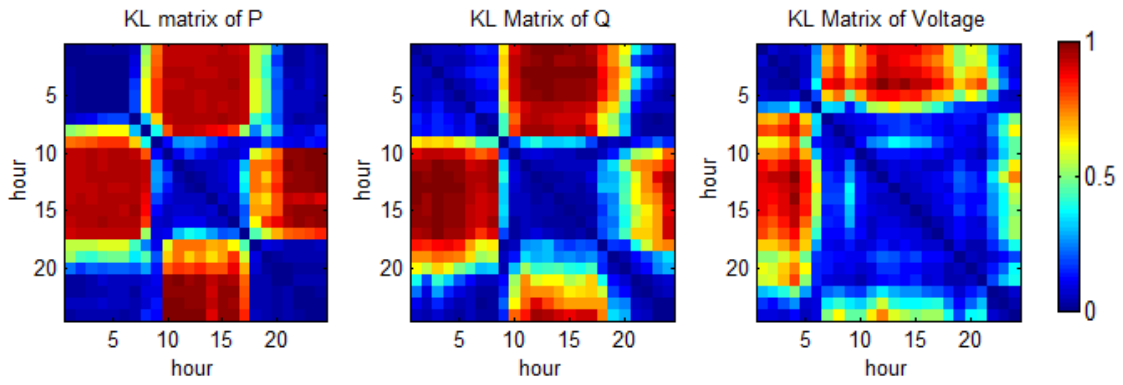

Figure 24. Normalized KL divergence matrices for real / reactive power, and voltage.

For the case shown in Figure 23, the daily 24 hours of the commercial load (Atlanta local time) are partitioned into working hours (red), off-working hours (blue), and hours in between (green). Since the load behaviors within working hours and off-working hours are highly consistent, these hours are merged. As a result, the number of models on the second layer of Figure 22 is reduced and the data for each time label increases correspondingly. Similarly, residential load and industrial load can be processed in the same way.

4.3.2 K-subspace Clustering

In practice, multiple load conditions can exist under the same load type and time label. As a result, on the third layer of the model, smart meter readings are further clustered into several load conditions so that each of the load conditions can be modeled by a static ZIP model.

Traditional K-means algorithm [71] clusters data based on their relative Euclidean distance to the nearest cluster center with an iterative process to adjust the centroid. The clusters' shapes are determined by the perpendicular lines between centroids. However, the smart meter readings of different load conditions are distributed in a very specific line-

shaped pattern close to each other as shown in Figure 25, which consistently leads to the failure to identify the correct clusters using traditional Euclidian-based k-means algorithm.

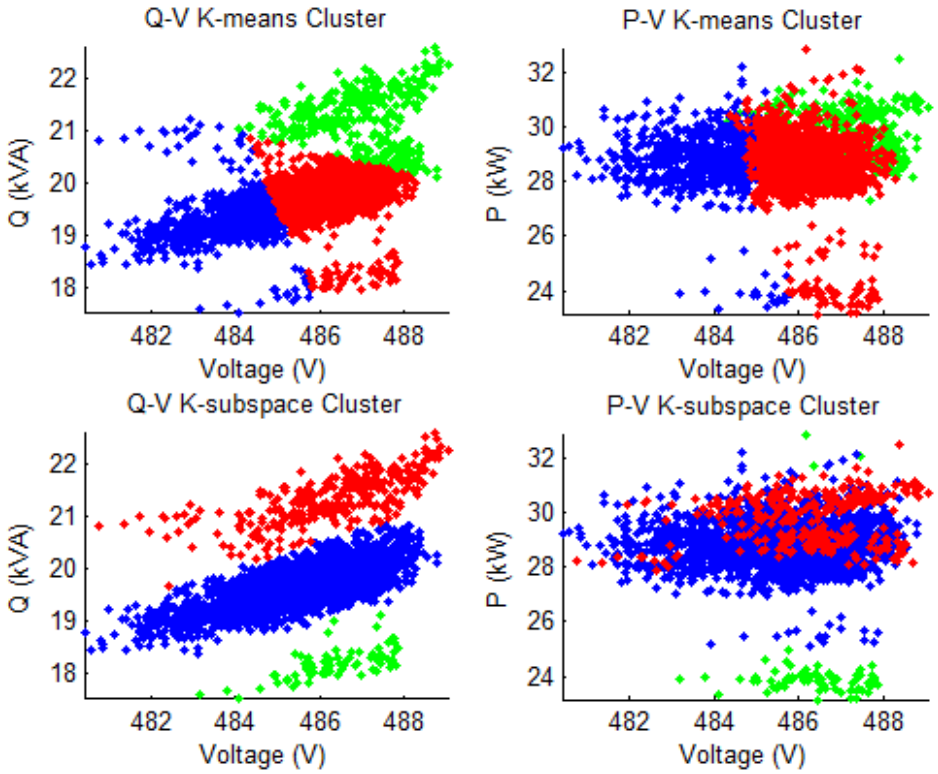

Figure 25. Comparison between K-subspace method and K-means method.

Given the line-shaped pattern we observe from the ZIP model, instead of using conventional k-means method, k-subspace method [72] is adopted which allows the detection and clustering of line-shaped data. The k-subspace method identifies line-shaped clusters by assigning each cluster C_k with a unit direction vector $\mathbf{a}_k$ and a center $\boldsymbol{c}_k$. The entire algorithm seeks to minimize the perpendicular distance of all the data points $\boldsymbol{x}_{k,i}$ to the line defined by $\mathbf{a}_k$ and $\boldsymbol{c}_k$ within each cluster, as shown in (25).

$$\min_{\mathbf{a}_k, \boldsymbol{c}_k} \sum_{i \in \mathrm{C}_k} Distance(\boldsymbol{x}_i, C_k) = \sum_{i \in \mathrm{C}_k} \|\boldsymbol{x}_i - \boldsymbol{c}_k - \alpha \mathbf{a}_k\| \tag{25}$$

where $\alpha = (\boldsymbol{x}_i - \boldsymbol{c}_k)^T \mathbf{a}_k$.

Figure 25 shows the Q-V and P-V plot of a commercial building during off-working hours on weekdays in the fall of 2012. Comparing Q-V plot with P-V plot, we can see that reactive power is more sensitive to voltage deviations than active power. As a result, the load conditions are clustered using only Q-V plot. In Figure 25, the clustering results are marked with different colors, where the cluster number k is set to be three.

The number of clusters or load conditions is determined through cluster evaluation. We first define the distance from cluster C_i to cluster C_j as (6). Similar to the KL divergence matrix, a cluster distance matrix can be formulated in the same manner. In the cluster distance matrix, small $Dist(C_i, C_j)$ and $Dist(C_j, C_i)$ indicates clusters C_i and C_j are very close to each other and should be merged. On the other hand, large $Dist(C_i, C_i)$ indicates a larger number of clusters k is required to identify all load conditions.

$$Dist(\, C_i, C_j) = |C_i|^{-1} \sum_{\boldsymbol{x}_i \in \mathrm{C}_i} Dist(\boldsymbol{x}_i, C_j) \tag{26}$$

where $\boldsymbol{x}_i$ is a member of cluster C_i.

Figure 26 shows how the number of clusters $k = 3$ is determined for the Q-V plot in Figure 25. A threshold h is set by experience to test the accuracy of k. In this case, h is set to be 0.1. The algorithm increases k until k equals to four when $Dist(C_1, C_2)$ and

$Dist(C_2, C_1)$ are both under the threshold h, which indicates the two clusters should be merged.

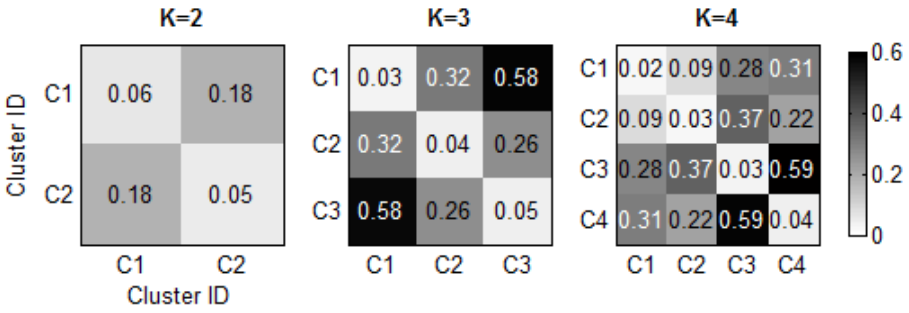

Figure 26. Cluster Distance Matrices with Different *k*.

After clustering, all smart meter data will be labeled by load type, time and load condition. Data with the same label is grouped to represent a single load condition. Then regression is performed to identify the parameter of the corresponding load condition model using (21-22).

4.4 Test Results Analysis

The data used in the study of this chapter comes from a historical smart meter measurement database collected by smart meters installed on the Georgia Tech campus. In order to enhance monitoring and reliability of the campus power network, smart meters were widely installed on Georgia Tech campus starting from in 2011. Currently, there are over 400 smart meters installed on the campus, covering each of the 200 buildings. Similar to most of the smart meters in the world, the data are recorded every 15 minutes including measurements of: real and reactive power (P, Q), power factor, voltage (V), and current for each phase. To illustrate the new modeling approach, different buildings are selected in this study, covering various load types such as commercial, residential and industrial loads.

4.4.1 Time Label Identification for Different Load Types

Various load types are studied to explore their differences in identifying the time label. In the study, a student residential hall and a family apartment are chosen as residential loads; an office building and a student center are chosen as commercial loads; and a chiller plant on campus is chosen as an industrial load. By using KL divergence matrices, their time label identification results for weekdays in fall are shown in Figure 27. Hours with consistent load behaviors are merged. Results shown in Figure 27 indicate that even under the same load type as residential load, different customers have their own power consumption pattern. For example, the peak hours of student residential hall and family apartment do not always overlap.

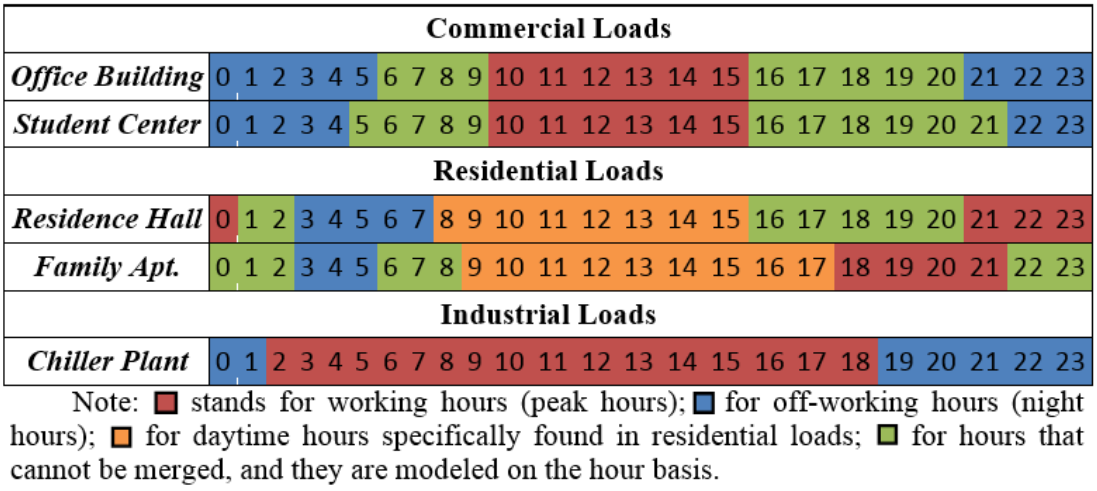

Figure 27. Time Label Identification Results (weekday, fall).

4.4.2 Data Mining-based Load Model

One of the key advantages of the data mining-based method is that a customized time-variant model can be built for every single customer equipped with a smart meter. To

illustrate the idea, we chose an office building on campus and built a time-variant load model using the proposed data-mining algorithm.

Once all of the smart meter data is clustered into the different times and load conditions, least square estimation is performed on each cluster to determine the parameters in (21-22). To suppress the noise from the data and avoid over fitting, power model is adopted and p_i and q_i are computed using Taylor expansion. Table 6 shows the partial regression results of an office building during the summer season.

Table 6 – Comparisons of Measurement-Based and Component-Based Method.

Office Building	Working Hours						Off-working Hours					
	$P(V)$			$Q(V)$			$P(V)$			$Q(V)$		
Cond.	p_1	p_2	p_3	q_1	q_2	q_3	p_1	p_2	p_3	q_1	q_2	q_3
1	4.71	-5.82	2.10	18.05	-29.57	12.52	0.42	0.704	-0.12	10.86	-16.53	6.67
2	28.00	-63.01	36.00	26.63	-45.44	19.81	-0.01	0.03	0.98	12.17	-18.88	7.71
3	--	--	--	--	--	--	0.018	0.999	-0.017	11.24	-17.21	6.97

Note: P_0 and Q_0 are adopted as the load real and reactive power at the nominal voltage.

4.5 Conclusion

A novel data mining method to build a time-variant load model is explored in this chapter, which takes advantages of the smart meter historical database.

The new modeling method offers several advantages with respect to existing methods. First, as long as smart meters are widely installed, there are no additional investments needed for advanced load modeling. There are no costly voltage stage tests,

no load component tests, no public surveys, and no validation tests (the model is based on real system measurements). Second, the new model is fully customized for every customer equipped with a smart meter, thus more accurate than other aggregated models. Third, the dynamically changing nature of the load is captured in the new model. Even the long term evolvements of load compositions are considered, which can be more significant when comparing current load composition with the ones in late 90s [73].

However, the proposed model has some limitations. Because reactive power is much more sensitive to voltage deviations than real power [74], more advanced data mining techniques are required to better capture the weak correlation between real power consumption and system voltage to improve the P-V model accuracy. Further studies may also include exploring the statistical information of the load conditions from the historical database and integrating the model into more advanced power system simulation and control applications.

CHAPTER 5. MACHINE-AIDED HOSTING CAPACITY ANALYSIS

The integration of distributed energy resources (DER), especially distributed solar photovoltaics (PV), has been gaining pace in the past decade. Solar PV is the fastest growing renewable energy source in the US [75]. Uncoordinated plug-in of PV systems may cause various issues in the distribution network, such as voltage spikes, thermal violations, and protection mechanism dysfunction. In Chapter 2, we study how to use stochastic models to detect unauthorized PV systems. In this chapter, we propose a machine-aided method to estimate PV hosting capacity of a feeder. The state-of-the-art hosting capacity analysis is quasi static time-series (QSTS) simulation, which is accurate but prohibitively expensive to run. The proposed method provides a strong support to speed up QSTS simulation thus making hosting capacity analysis a much easier task with very fast processing speed and low memory consumption.

5.1 Hosting Capacity Analysis and Quasi-static Time Series Simulation

5.1.1 Hosting Capacity Analysis

Hosting capacity analysis is a comprehensive study to answer the question of how many renewables can be connected to a distribution feeder. In many U.S. states, customers need to obtain a permit from a permit agency before installing any PV system. The major reason for requiring a PV system permit is to prevent uncoordinated PV interconnection, which may jeopardize grid reliability and power quality. The evaluation of a PV system interconnection application calls for a comprehensive hosting capacity analysis on the

specific feeder. However, due to the volume of the PV permit applications, currently, the permit agency may not have enough time and resources to run hosting capacity analysis for all applications. Instead, rough rules are established to determine whether a permit can be issued. For example, to clear the permit application queue faster and keep up with the incoming applications, a permit agency may issue the permits merely based on the PV system sizes.

One obvious problem with the rule-guided PV permit issuing policy is the over conservative evaluation result. Since the true PV hosting capacity is related to numerous factors, a rule-guided policy usually sets the PV plug-in bar as high as possible to make sure no significant impact will be introduced. A better way to evaluate a PV plug-in application is to run hosting capacity analysis, which provides an accurate and thorough examination of the potential impacts of a proposed PV system. However, three major barriers need to be conquered before hosting capacity analysis can be widely used and incorporated into the utility's business model.

First, the hosting capacity analysis requires a calibration of the feeder model. If the distribution system is not well modeled or the topology of the feeder is unknown, it is very difficult to perform hosting capacity analysis. Thanks to widely installed smart meters, many researchers have proposed various algorithms to calibrate both the topology and parameters of the distribution network using smart meter measurements [76]-[77].

Second, hosting capacity depends on many factors, which makes it very difficult to generate a set of rules or guidelines that are applicable to all feeders. To determine the hosting capacity, we need to know the location of the PV system, the feeder topology, the

local solar irradiance, and the control strategy of the PV inverters. When it comes to determining the hosting capacity, there is not a necessary correlation or a rank of importance among these factors. As a result, without running hosting capacity analysis, it is difficult to guarantee the safety of adding a new PV system. .

Third, the PV installations may also cause some unexpected problems that only time-series simulation can detect. One of the most significant issues is the compatibility between the PV system and system controllers. Distribution controllers, such as regulators and capacitors, are meant to keep the feeder voltage within a normal range. When the system load changes, system controllers will take action to maintain the feeder voltage. The system controller's action delay is designed to avoid oscillations when the load changes around the control action boundary. PV systems have high variability. Despite the control delays, frequent PV output spikes might trigger unnecessary capacitor and regulator oscillations. These unnecessary oscillations will seriously shorten the life span of these controllable elements.

5.1.2 Scenario-based Hosting Capacity Analysis

Since the hosting capacity of a feeder depends on numerous factors, the most intuitive solution to conducting a comprehensive hosting capacity analysis is solving the power flow with several distinct scenarios. This hosting capacity analysis usually takes as many distinct scenarios as possible in order to cover the combinations of all these factors. Since this approach seeks to summarize the hosting capacity of a feeder by integrating various scenarios, we categorize it as scenario-based hosting capacity analysis.

The scenario-based approach is rooted in a conservative philosophy of designing a system that can withstand the worst possible scenario test. This philosophy is very common in the transmission grid analysis due to high reliability requirements. The scenario-based approach seeks to capture the worst possible scenarios, such as installing the PV system at the worst location and choosing a load level that most likely to cause over voltage. However, worst-case scenario analysis is not the best strategy for distribution system hosting capacity analysis.

First, the scenario-based analysis results depend strongly on how the researchers choose and design the scenarios. One can always come up with a more extreme case that will invalidate the previous hosting capacity result. Compared with the transmission network, distribution utilities might be able to tolerate a few hours of over voltage at a few buses.

Second, scenario-based analysis cannot provide a thorough exam of how the circuit will behave in the long run, e.g. one year. A scenario-based algorithm can answer the hosting capacities of the circuit under the predesigned scenarios. But it cannot answer the questions on how often the scenarios are likely to occur. If the scenario-based analysis reports a slight voltage violation, we do not know how likely this scenario is and how long the violation will be. If the violation scenario only lasts a few hours per year, we are getting a very conservative hosting capacity result by reducing the hosting capacity to accommodate a rare scenario.

5.1.3 Quasi-static Time-Series (QSTS) Simulation.

To overcome the weakness of the scenario-based, quasi static time-series (QSTS) simulation is proposed. For a given circuit, the inputs of the QSTS simulation include load and PV profiles, regulator and capacitor's control logics, and the topology and parameters of the circuit itself. The QSTS simulation takes the time-series inputs and solves power flow on every time unit. The QSTS simulation also keeps track of the control logic of system controllers including the control action boundary and delays during the simulation. The outputs of the QSTS simulation records all the behaviors of the simulated circuit including the state of system controllers and bus voltages through time. A general flow chart of the QSTS simulation is shown in Figure 28, where the simulation time resolution is 1 and the simulation time is $SimTime$.

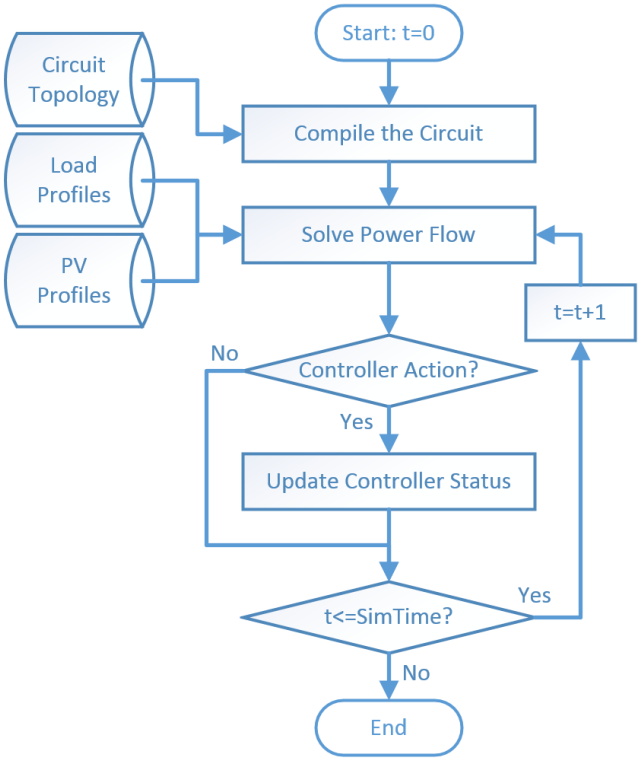

Figure 28. The QSTS simulation flow chart.

In order to better describe the differences between QSTS simulation and scenario-based methods, let us consider an analogy from a musical rehearsal, where the system violations are the bad acts in the show. Compared with scenario-based methods, the QSTS simulation rehearses all the chapters of the script line by line from the very beginning to the end. During the rehearsal, the QSTS simulation also uses a camera to record all the details of the show through time. Thus, after the QSTS simulation, all details of the musical can be reconstructed by examining the video recording. We can easily count the number and durations of bad acts, analyze their cause, and propose corrective actions. Instead of rehearsing the whole script through time, scenario-based methods take a different approach by picking up a few acts per chapter and checking whether the picked act is a bad act. This will inevitably miss many errors during the play.

According to the previous analogy, the advantage of the QSTS simulation is obvious, it offers much more information compared to scenario-based methods. The utility can examine bus voltages, thermal loading, and regulator tap positions for any arbitrary point in time. However, the disadvantage of QSTS simulation is the computational time required. Scenario-based methods run much faster, since only a few scenarios are constructed and analyzed. According to [78], yearlong high time resolution QSTS simulation can take from 10 to 120 hours to run for realistic feeders. This is also the only reason, why the QSTS simulation is not commonly used by the industry in circuit hosting capacity analysis.

One might propose to run a lower time resolution QSTS simulation to cut down the computational time. For example, instead of running QSTS simulation second by second, one can run the same simulation hour by hour. However, according to [79], yearlong high-resolution QSTS analysis is necessary to model the impacts of various distributed resources

on system controllers' behaviors. This is because most system controllers in the distribution system have various delays from a few seconds to several minutes. A 5-second or higher time resolution QSTS simulation is necessary to capture of potential regulator/capacitor behaviors, including potential controller oscillations. Figure 29 shows an episode of system regulator and capacitor oscillation caused by high PV penetration. In this specific system, the control delay of the regulator is 15 seconds and the delay for the capacitor is 30 seconds. If we take a lower time resolution, for example an hour, the QSTS simulation will not be able to capture these system controller oscillations.

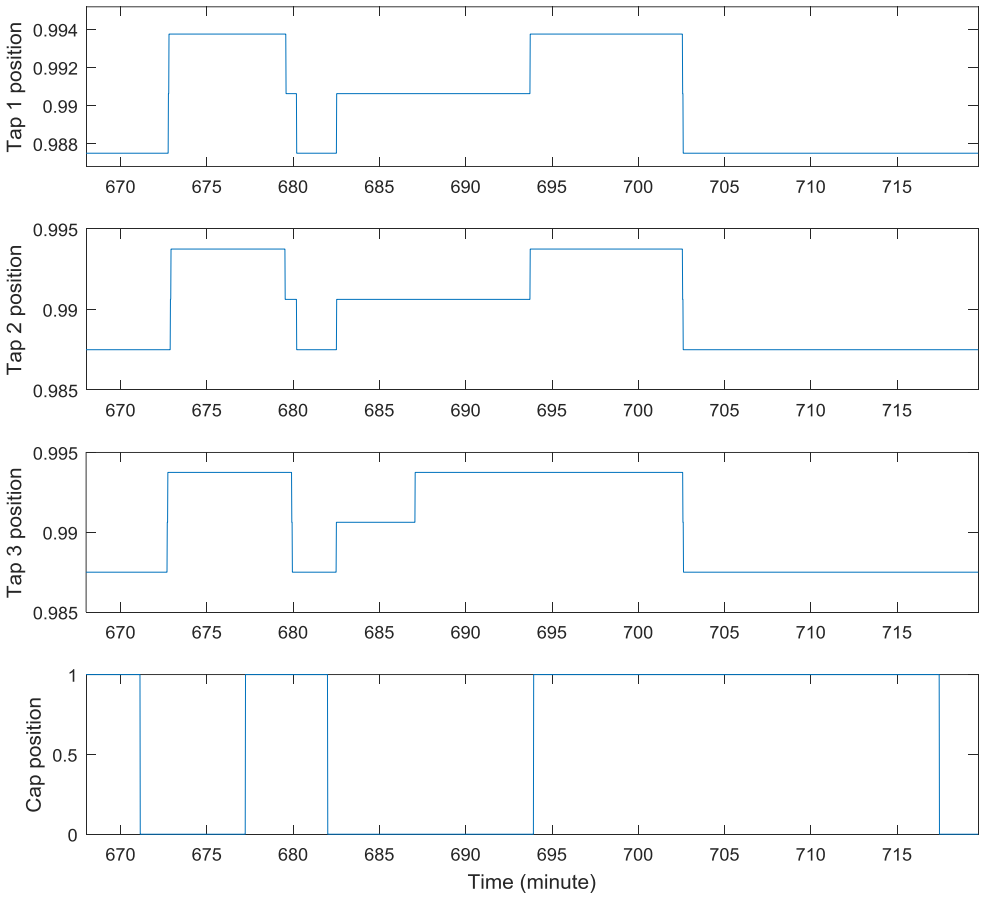

Figure 29. System controller oscillations.

As a result, in the face of numerous distributed resources and system controllers, it is critical to develop a fast approach to run QSTS simulation to replace the scenario-based hosting capacity analysis.

5.2 A Machine Learning Solution Formulation

The objective of QSTS simulation is to evaluate the hosting capacity through various indices, such as over/under voltage duration per year, number of regulator and capacitor actions per year, hours with feeder thermal limit violations, and so on. QSTS simulation provides the accurate values of these indices to support the hosting capacity analysis. Under this QSTS simulation setting, the inputs to the simulation are: load and PV profiles, network topology, and system and controller parameters; and the outputs of the simulation are the abovementioned annual indices, as shown in Figure 30.

According to Figure 30, if the brute force QSTS simulation method is used, the computation time for the QSTS simulation increase linearly with the duration of the simulation. Thus, the brute force algorithm has the time complexity of $O(n)$, where n stands for the number of time points in the simulation. The lower bound for the total computation time is $n\Delta t$, where Δt is the computation time required to solve one power flow.

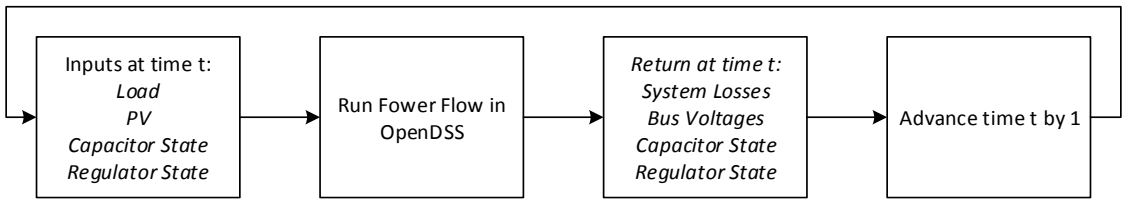

Figure 30. Brute force QSTS simulation flow chart.

In this section, we formulate the fast QSTS simulation task as a machine learning process. Machine learning can establish a direct mapping between the inputs and outputs using a constant length of QSTS results as training data. Thus, the training time of machine learning approaches can be constant. Moreover, with a proper choice of machine learning algorithm, the computational time or testing time is negligible. Machine learning approaches seek patterns between the QSTS simulation inputs and outputs. The learned patterns will allow machine learning models to generate outputs that are similar to brute force QSTS simulation given the same inputs. If successful, we can safely bypass brute force QSTS simulation and use a machine learning model as an alternative path to acquiring QSTS simulation results, as shown in Figure 31.

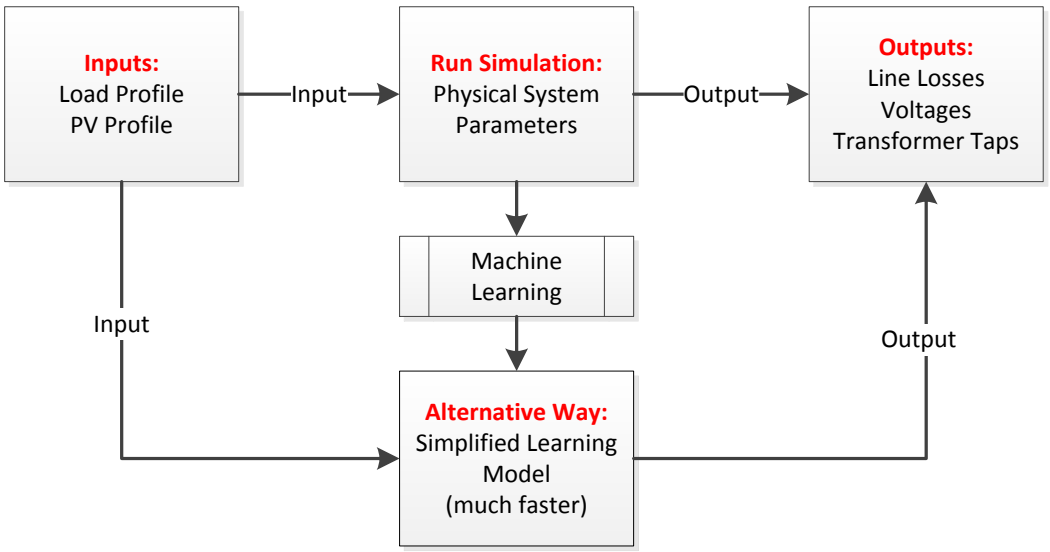

Figure 31. Machine learning problem formulation.

Apart from fast computational speed, there are several other advantages of using machine learning models to speed up the QSTS simulation. One of the advantages is that the machine learning model can provide additional insight into the operation of the grid by

interpreting the mechanisms and conditions under which operational violations occur, thus providing meaningful insights for future decision making. For example, if a linear model is used to predict the regulator tap-change frequency, a relatively large positive coefficient for PV output will imply that the PV output has a very strong influence on the regular tap-change frequency for the given feeder. Thus, reducing the PV size is very likely to significantly reduce the wear and tear on the regulator.

Another advantage of machine learning models is that the model complexity does not depend on the number of system buses. The brute force QSTS simulation complexity grows linearly with the number of system buses. This is because the power flow solving time is approximately linear with respect to the number of system buses. However, the complexity of machine learning models only grows when we increase the number of model features. For example, if we use a regression model to predict system losses, the features of the regression model can be system load profiles, PV system output profiles and so forth. As long as we are using the same features, the complexity of the regression model stays the same.

5.2.1 Model Evaluation and Selection

The power of machine learning lies in its ability to generate simple mathematical models to replace true physical systems that are too complicated to calibrate. Since there are numerous machine learning methods that can be applied to speeding up QSTS simulations, it is necessary to establish a guideline for choosing the most accurate and efficient algorithms for our purpose. We use the following general guidelines:

First, we prefer low complexity algorithms that can be trained within a short period of time. Since the overall objective of the study is to speed up the QSTS simulation, the machine learning algorithm itself must be more efficient than solving power flows with the brute force method. Thus, computational speed is one of the most important factors in model selection. For example, under the same accuracy criteria, we prefer linear regression to random forest.

Second, we prefer algorithms with parameters that can be used to extrapolate knowledge of the physical system. Although many of the machine learning algorithms contain some characteristics of the original physical system, they are not the same. For example, logistic regression is better than support vector machine. Parameters in logistic regression models have a physical meaning of logit, while support vector machine methods project the original problem into higher dimensional space whose physical meaning is unclear.

Third, we can tolerate models with higher variance, as long as they have low bias. For a high variance but low variance model, we might not have a very accurate predictor of the QSTS results for a specific hour of the year, but the annual aggregated indices prediction is very accurate. The hosting capacity analysis relies more on the annual aggregated indices.

In this research, we first evaluate general machine learning algorithms that are not specifically designed for the application of fast QSTS simulation. These general algorithms include unsupervised learning such as clustering and supervised learning such as regression. Then, we discuss the challenges of speeding up QSTS simulation which

explains why general machine learning algorithms have difficulty achieving both fast computational speed and accuracy. After a thorough analysis and discussion of the challenges, a more advanced machine aided approach is introduced, which is specifically designed for the purpose of fast QSTS simulation, and which satisfies both the speed and accuracy requirements. The proposed approach involves a plane-based model that takes advantage of the geometric insight of distribution system voltages given the inputs of PV and load, and considerations of the controller operations.

5.2.2 Unsupervised Learning Approaches

Before introducing the plane-based model, a few general machine learning algorithms are explored including unsupervised and supervised learning algorithms. Commonly used unsupervised learning (clustering) algorithms include k-means, hierarchical clustering, spectral clustering, density estimation, and so forth. These algorithms all serve to group similar data points into clusters.

Under the unsupervised learning setting, we cluster the QSTS input into several clusters and run QSTS simulation on the representative samples of these clusters. The QSTS simulation results of the samples are later used to reconstruct the results for the whole clusters. The reason behind the clustering algorithm is that similar inputs of the QSTS simulation is likely to generate similar outputs. For example, if we pick two distinct days of the year that share very similar load curve and PV output curve, we may assume that the QSTS simulation results of the two days would be similar (same number of controller actions, same period of over voltage, and so on). Following this insight, we develop an unsupervised learning algorithm that follows these steps:

1. Cluster the daily inputs (PV and Load) into k daytime and night-time clusters.
2. For each cluster, we sample the load and PV profiles through bootstrap sampling without replacement. The sample size of each cluster is linearly correlated to the size of the cluster.
3. We run QSTS on the samples from each cluster.
4. We use the QSTS simulation results of the sampled data to reconstruct the annual QSTS simulation results.

In Figure 32, we cluster the net load profiles into 10 night-time and day-time clusters. The centroids of the 10 clusters are plotted in different colors. The daytime net load centroids have much more noise compared to the night-time net load centroid. This is because the PV output during the daytime creates large variations in net load.

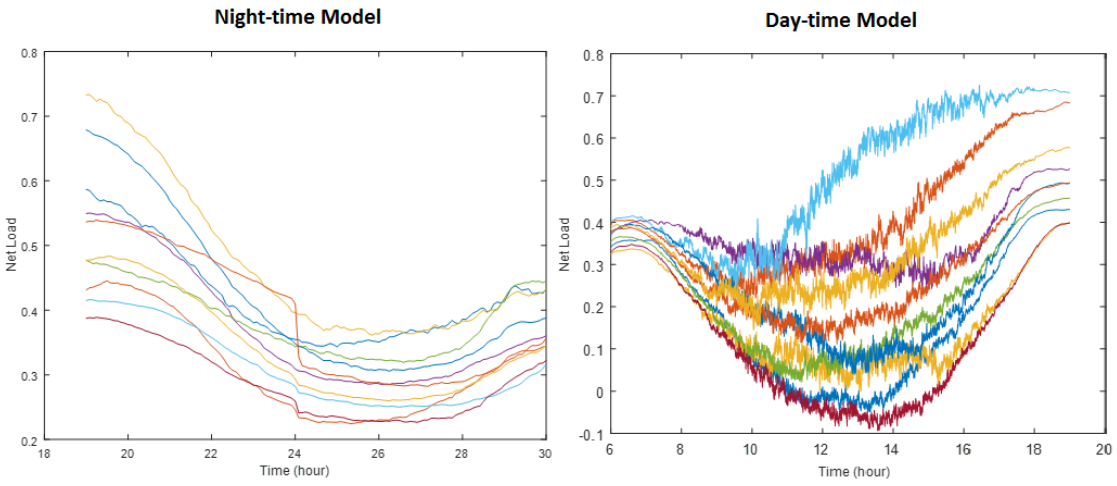

Figure 32. Centroids of the net load clusters.

In the next step, we sample the load and PV load curves from each of the 10 clusters using bootstrap sampling without replacement. The number of samples drawn from each cluster is linearly correlated with the cluster size. Then, we run the QSTS simulation on

these samples and collect all relevant indices. The QSTS results for the samples are later used to reconstruct the annual QSTS results. For a given cluster with n members, we draw m samples from the cluster, a sampling rate of m/n. The QSTS results of the all n members in the cluster are derived by taking the mean value of the m sample QSTS results.

The unsupervised algorithm has many sources for randomness which adds variance to the final estimated annual QSTS results. First, depending on the unsupervised clustering algorithm, the load/PV profile clustering results differ slightly on different trials. For example, k-means algorithm may converge differently depending on the deployment of initial cluster centroids.

Second, the bootstrap sampling is a random process. Although the bootstrap sampling algorithm is an unbiased estimation method, the brute force QSTS simulation is very slow to run and cannot yield a high sample/population ratio to guarantee an accurate estimation of the total population. Figure 33 shows the boxplot [44] of the number of tap actions for the ten clusters considered in Figure 32. The central red mark is the median, the edges of the box are the first and third quartiles, and the red cross stands for outliers. From the boxplot, we can see that the numbers of tap actions of load curves are similar within a cluster but vary across clusters.

Third, the initial system controller status also plays an important role when running QSTS simulation on sampled load curves. Since we break down the load/PV profiles into daytime and night-time segments, the continuation of system controller status across the breaking points is lost. This may create a small discrepancy at the beginning of each QSTS simulation for each sampled load profiles.

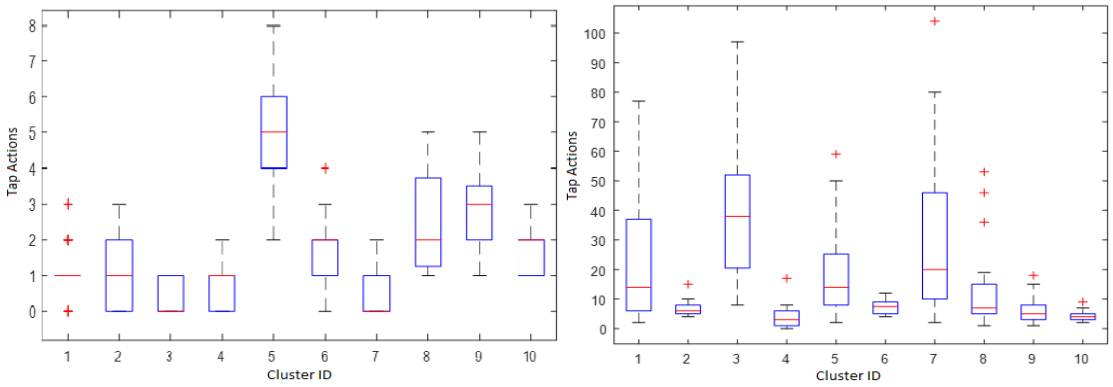

Figure 33. Boxplot of tap actions for each cluster.

In order to further illustrate the randomness embedded in the unsupervised learning algorithm, we quantify the randomness of the algorithm by running the algorithm multiple times as shown in Figure 34.

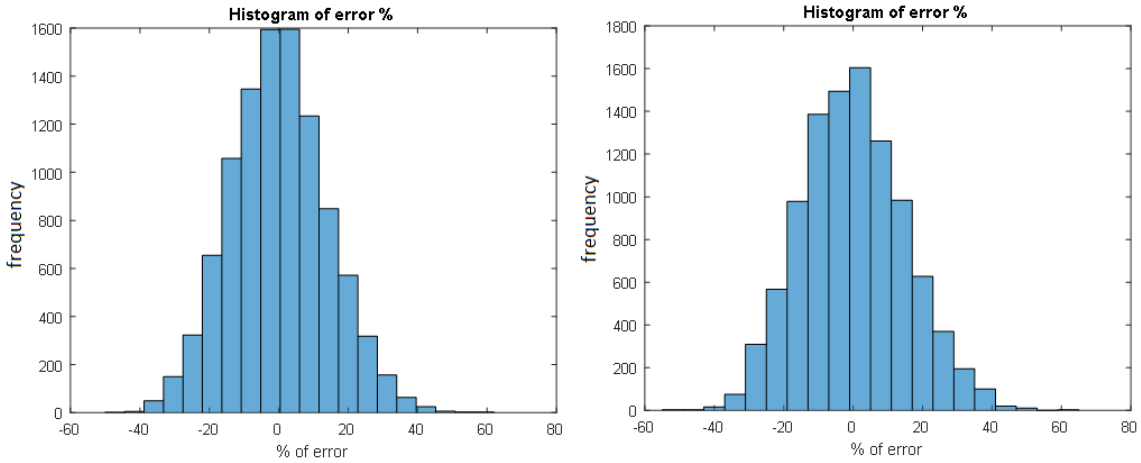

Figure 34. Stability study for night-time and day-time models.

Figure 34 shows the relative error of the reconstructed annual results after 10,000 bootstrap trials. The model performs well for night-time data but bad for day-time data. This is because the high variability of PV outputs is not fully captured by the k-means clustering algorithm. Thus, the proposed unsupervised learning algorithm does not

guarantee results accuracy all the time. We run 10,000 experiments and acquire the following statistics:

- Night-time
  - 50% of the time the result falls in the 5% error band
  - 81% of the time the result falls in the 10% error band
- Daytime
  - 31% of the time the result falls in the 5% error band
  - 56% of the time the result falls in the 10% error band

The advantage of using an unsupervised learning algorithm to solve QSTS is that the algorithm does not require training data. We can implement clustering algorithms before running QSTS simulations. However, the algorithm has poor accuracy, not sufficient to support hosting capacity analysis.

5.2.3 Supervised Learning Approaches

The other major category of machine learning is supervised learning methods, which are suitable for classification and regression problems. For simplicity, let us first focus on the annual numbers of regulator and capacitor actions as the index of interest. Instead of using a classification model, a regression model is more powerful and easier to implement for this specific application. This is because system controller status is better modeled as an integer than a label.

The first step in running a regression algorithm is feature selection. In this study, we take daily features as regression inputs, aiming at creating a short cut between the time series simulation inputs and outputs as shown in Figure 31. These features occur daily, such as daily mean energy consumption, daily maximum net load, and daily variance of PV system output. During the feature selection process, we first generate as many features

as possible. All the input predictors can be categorized into two categories: load features and PV features. Then, we select the most statistical significant features from the feature pool. The features used in the regression model includes daily minimal/maximum load, daily average load, daily average PV output, variation score of the load, and variation score of the PV output.

The second step is to select a suitable regression model. There are a wide variety of regression algorithms including linear regression, LASSO, regression tree, and ensemble learning algorithms. However, the capability of the regression model is usually proportional to the complexity of the model, and more complicated model requires more training data. To achieve a reduction of 80% of the simulation computational time, , we can at most run the brute force QSTS simulation for 72 days, which is equivalent to 72 data points as training data. This will rule out complicated regression algorithms such as ensemble learning algorithms and neural networks. Thus, linear regression is more suitable for prediction of system annual controller action numbers.

In a distribution system with limited PV penetration, a linear regression model works well, approximating well the true behaviors of the power system. The estimated annual tap action vs. the true annual tap action is 7034 vs. 7202. Since the PV output is sufficiently small, the system controller actions are mainly driven by the load profile, as shown in Figure 35.

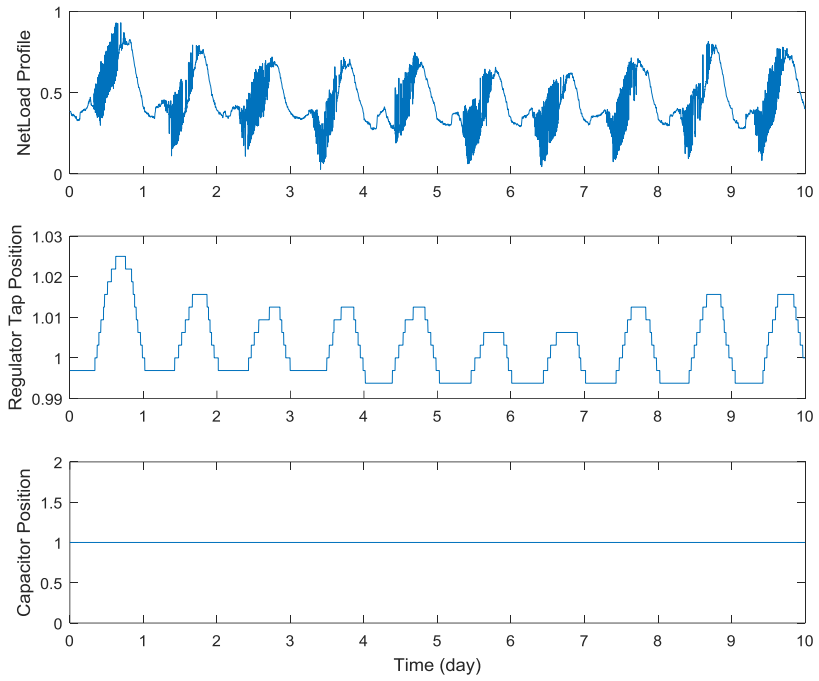

Figure 35. QSTS simulation results for a system with 10% PV penetration.

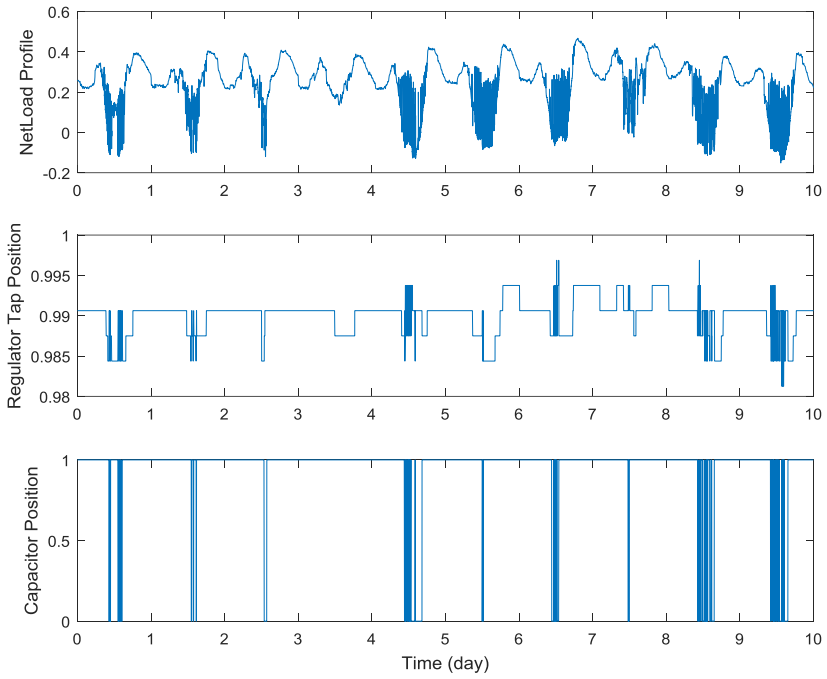

Figure 36. QSTS simulation results for a system with 40% PV penetration.

However, for systems with significant PV penetration, the linear model no longer works. A much larger PV size may constantly create negative net load as shown in Figure 36. Under this scenario, most of the controller actions are driven by the PV output especially when the system load level is low. The large variation of the PV output can lead to frequent capacitor actions, which are usually followed by a series of regulator tap oscillations, as shown in Figure 29. An alternative solution is to introduce nonlinearity to the model to better capture the complicated interactions between the system controllers and renewable energy resources. In this study, a decision tree is established. The decision tree uses the same features as linear regression to cluster the input into different groups to reduce the variance of the target variable under the same leaf. Each leaf of the tree is a linear regression model. In this study we build a tree with two branches based on the daily mean load value. Different linear regression models are built for days with different mean load values. The proposed model reduces the prediction error to around 10%.

5.3 Challenges in Speeding up QSTS Simulation

Although many general machine learning algorithms may serve to improve the QSTS simulation speed, these algorithms have limitations. On the one hand, it is very difficult for general algorithms to produce an unbiased result with a variance that is small enough. On the other hand, these general algorithms call for a relatively large training data set and a balance of accuracy and efficiency is difficult to achieve.

In fact, the QSTS simulation is a highly specialized problem. Most of the general machine learning black boxes will have difficulty learning the complicated patterns. This is because the distribution system state is a highly discrete function and time correlation is

embedded along the QSTS simulation. This is further explained in the three challenges of speeding up QSTS.

5.3.1 Multiple Valid Solutions Challenge

The first challenge in speeding up QSTS simulation is multiple valid solutions. For a given QSTS input, multiple valid solutions mean that the system could have multiple valid power flow solutions depending on different system controller states. To better illustrate this , we take the system regulator tap control as an example.

A regulator tap aims to maintain the system bus voltage within a normal range. Let $V_{regCtrl}$ denote the input control voltage of a regulator. The regulator control keeps $V_{regCtrl}$ within a voltage band (V_{regMin}, V_{regMax}) by changing the tap position accordingly. When $V_{regCtrl}$ moves above V_{regMax}, the regulator control will trigger a tap switch event to move the tap to a lower position; similarly, when $V_{regCtrl}$ drops lower than V_{regMin}, regulator will trigger a tap switch event to move the tap to higher position.

We introduce a graphical representation of the regulator control strategies, as shown in Figure 37. $V_{regCtrl}$ is not linearly correlated with the load. However, in most distribution systems, the error introduced by linearizing the correlation between $V_{regCtrl}$ and load can be neglected. The linear sensitivity assumption is discussed in detailed later in the 5.4.1 section, where we assume a linear correlation between $V_{regCtrl}$ and the system load.

In Figure 37, we model each tap position as a solid line. When the load increases from L_0 to L_1, $V_{regCtrl}$ will drop from V_0 to V_1; similarly, when the load decreases from L_0 to L_2, $V_{regCtrl}$ will increase from V_0 to V_2. As long as the load maintains within L_3 and L_4,

no tap event will be triggered. However, when the system load moves below L_4 or above L_3, $V_{regCtrl}$ will move outside the voltage band between V_{regMin} and V_{regMax}. This will trigger a tap action, where the tap will move to the adjacent tap position, which corresponds to the adjacent lines in the graphic model.

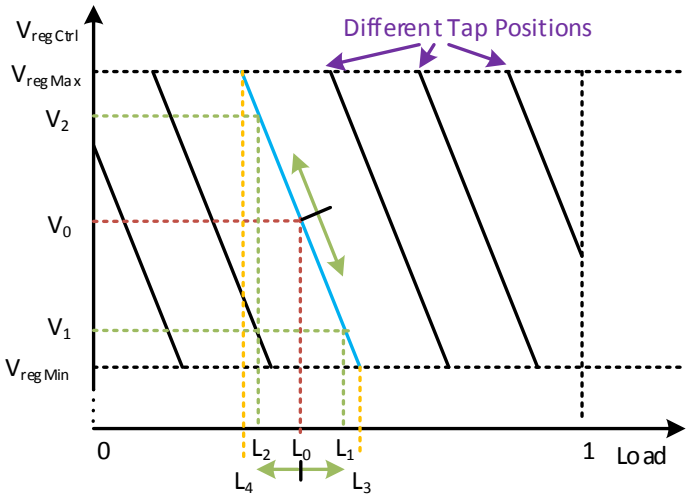

Figure 37. Regulator control input voltage vs. system load.

Given the proposed graphic model, we can easily explain how the regulator control could create multiple valid power flow solutions under the same load, as shown in Figure 38. When the load equals to L_1, then there is only one possible tap position. However, when the load equals to L_2, there are two possible tap positions. The two valid tap positions will lead to two different $V_{regCtrl}$ as V_2 and V_3, both of which are within the voltage band of the regulator's control logic.

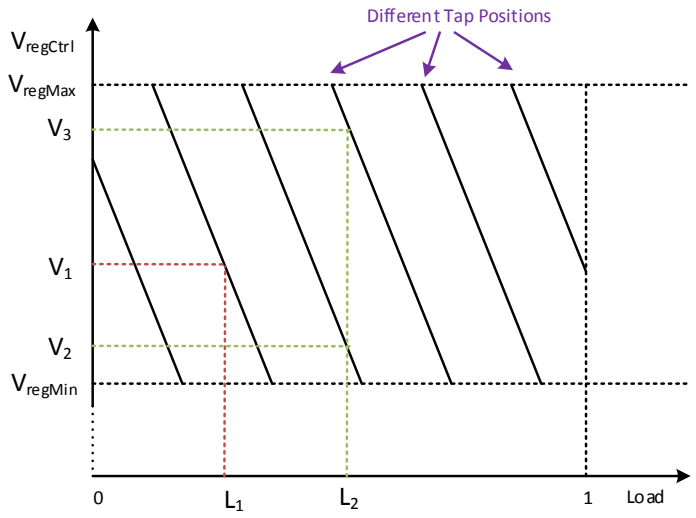

Figure 38. Multiple valid solution caused by regulator control settings.

In practice, it is required that, under any load level, a regulator must have at least three distinct valid tap positions. Thus, a realistic regulator setting with the "three-overlapping tap" rule will have a graphic representation as shown in Figure 39. In other words, for any given load level, a regulator will yield at least three valid tap positions, each resulting in a distinct power flow solution.

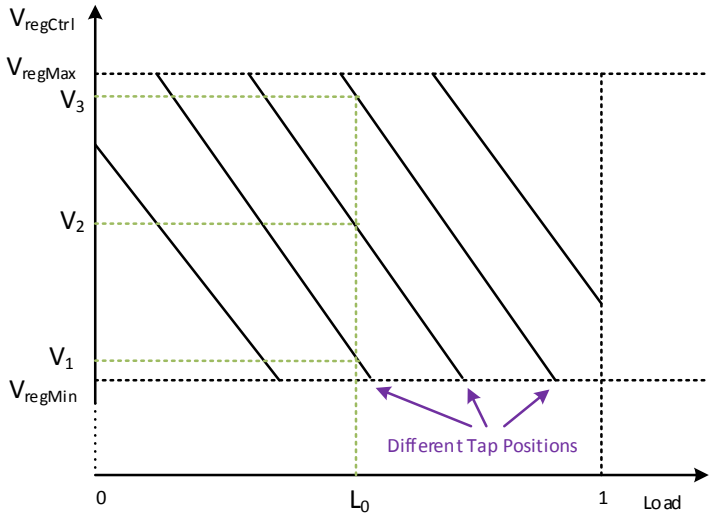

Figure 39. The "three-overlapping tap" rule of the regulator control setting.

The multiple valid solutions challenge emerges due to the design of system controller settings. The specific design can effectively prevent regulator tap switching oscillations. However, it creates a great challenge for machine learning algorithms, which seek to establish a one to one mapping between the QSTS inputs and outputs. In the face of the multiple valid solutions, the one to one mapping no longer exists. The same load inputs will yield multiple possible voltage solutions. The most intuitive solution to this is introducing time dependency and time correlation.

5.3.2 Time Dependency and Time Correlation Challenge

According to the previous section, a static machine learning model that only takes the inputs at time instance t is unable to predict the outputs because of the multiple valid solutions challenge. To eliminate this uncertainty, it is necessary to introduce time dependency to machine learning models. In other words, this requires the machine model to incorporate some information from the previous time instances.

In the brute force QSTS simulation, the time dependency and time correlation is naturally incorporated in the system controllers' control logic when the simulation time advances second by second. By studying the brute force QSTS simulation, we learn that given the previous system status, the current power flow solution can be uniquely identified.

This idea is further illustrated by an example, as shown in Figure 40. In this example, we assume the load at time instance t is L_1. If we do not know the previous system states, there are two possible tap positions under the load level of L_1, corresponding to two distinct controller input voltages V_2 and V_3. However, if additional information is given regarding

the system status at previous time instances, we can eliminate this uncertainty. For example, let us assume the system status is operating at point E previous to time instance t. When the load decreases from L_2 to L_1, the system operating point moves from E to F and we expect no tap switch action. Under this scenario, it is not possible to trigger a tap action and consequently, the power flow solution can be found. Similarly, on the other case, let us assume the system is operating at point A with load equals to L_0 previous to time instance t. We can expect the system to operate at point D when the controller input voltage equals V_2, as the load increases from L_0 to L_1. Under the second scenario, the system operating point will first move from point A to point B. A tap action is triggered at point B as the load continues to increase. After the tap switch action, the system operation lands at an adjacent tap position at point C. As the load continues to grow, the system operating point will further move from point C to point D. Thus, incorporating time dependency can eliminate the multiple valid solutions challenge.

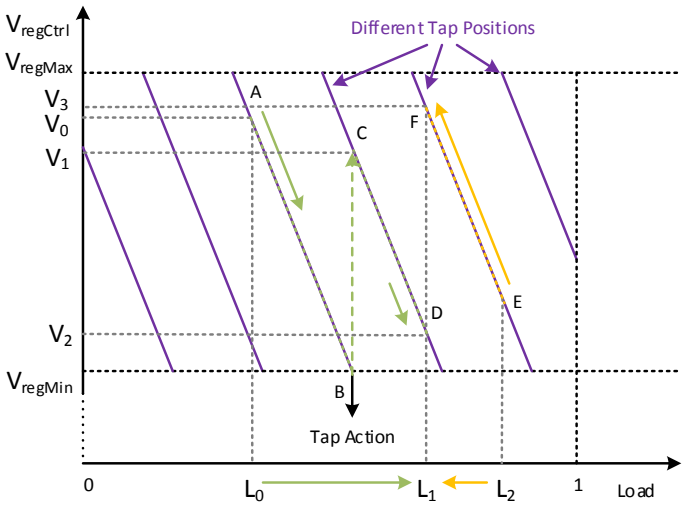

Figure 40. The time dependency of QSTS simulations.

Except for the continuous state shifting of system controllers, the control delay embedded in the control logic also plays a very important role in the time dependency challenge. In the previous example, we did not consider the delay of system controllers. In practice, to protect system controllers from oscillations, a delay is set whenever the controller's input voltage moves out of the control boundary. This can be shown in Figure 41, where the load increases from L_0 to L_1 monotonically. In Figure 40, a tap switch is triggered instantly at point B when $V_{regCtrl}$ drops below V_{regMin}. In Figure 41, there is no immediate tap switch action at point B, instead, a timer is initiated with a delay time of d seconds. Once the timer is initiated, if condition $V_{regCtrl}<V_{regMin}$ continues to hold for the next d seconds, a tap action will be triggered when the delay time is over. In Figure 41, when the load increase from L_0 to L_2, no tap action is triggered, but a timer is initiated. As the load continues to grow, the tap switch action is triggered at C, when the delay time is over. However, if the load falls back and $V_{regCtrl}$ becomes greater than V_{regMin} before the delay time is over, no tap action will occur. The timer is reset once $V_{regCtrl}$ falls back into the control boundary. Although the system will experience a brief duration of $V_{regCtrl}$ voltage violation because of the delay, the control delay can help eliminate unnecessary oscillations and significantly increase the life span of system controllers. For example, in Figure 41, if the load jumps from L_0 to L_3 and falls back to L_0 in 3 seconds, the three-seconds load spike will cause two tap switch actions without the delay.

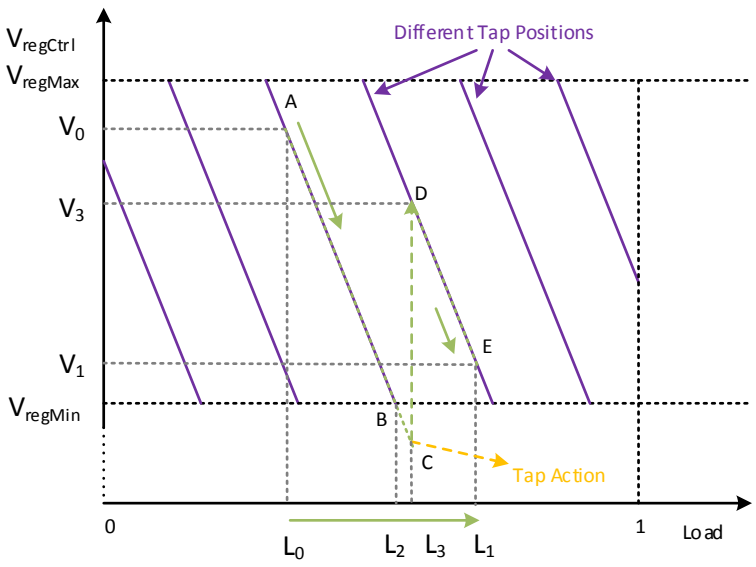

Figure 41. The delays of system controllers.

Since both the previous system controller states and the delay of system controllers are necessary to eliminate multiple valid power flow solutions, an effective machine learning model must be modified and incorporate time dependency. However, enhancing a machine learning model usually involves a trade-off between model complexity and model accuracy, which is another challenge discussed in the next section.

5.3.3 Model Complexity and Accuracy Trade-off

Both the abovementioned supervised and unsupervised machine learning models are static models, which take the time series QSTS inputs and produce the outputs in batches, as shown in Figure 42. In other words, these algorithms predict the QSTS simulation results without referring to the results from the previous time stamps. On the one hand, this static model will allow for a much faster computational speed. On the other hand, the fast speed comes at the expense of ignoring time dependency of QSTS simulation, which results in

low model accuracy. In other words, for a given QSTS input, static models can only return a statistically most likely solution from multiple valid solutions.

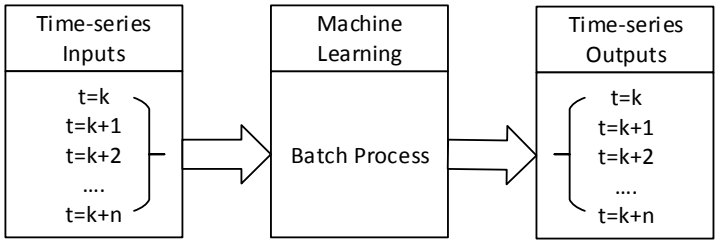

Figure 42. Static machine learning models with batch process.

To incorporate time dependency of the QSTS simulation, static machine learning models must be transformed to dynamic models that take the QSTS inputs in time sequence and advance the simulation second by second just as in a brute force QSTS simulation. As shown in Figure 43, the machine learning model not only takes the current QSTS inputs but also previous system controller states to predict the current system response.

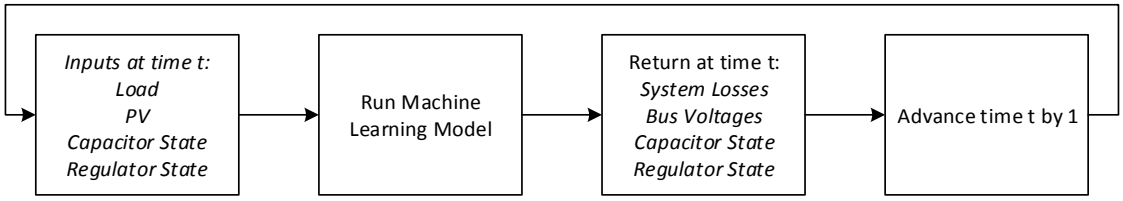

Figure 43. Dynamic machine learning models with batch process.

The dynamic model frame as shown in Figure 43 is very similar to the brute force QSTS simulation method shown in Figure 30. The major reason for the slow speed of brute force QSTS simulation is the repeated call to solve power flow at every simulation time unit. Although each power flow solution only takes a fraction of a second, the total computational time of one-year-one-second-resolution brute force QSTS is the single

power flow solution time times 31.5 million solutions. This easily scales up to hundreds of hours for a practical distribution network with thousands of nodes. Similarly, since a dynamic machine learning model also advances second by second, the total computational time can be very expensive.

Per the above discussion, due to the complexity of the QSTS simulation task, it is very difficult to maintain a balance between model accuracy and complexity, as shown in Figure 44. Thus, we need to revisit all the machine learning models discussed so far and seek alternative approaches.

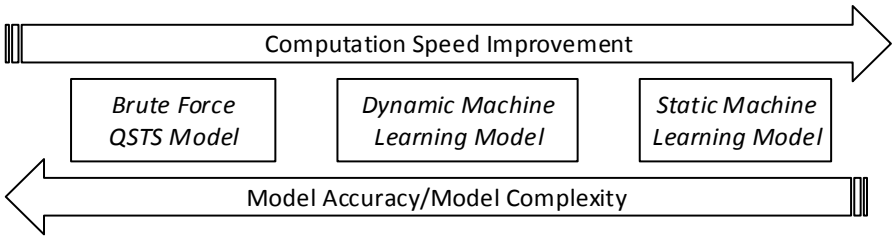

Figure 44. Model complexity and efficiency trade-off.

5.4 Plane-based Machine Learning Model

Most of the machine learning models are just mathematical representations of the true physical system. A machine learning model seeks to provide a similar input-output mapping as the true physical system.

From a broader point of view, the power flow solver model is also a mathematical representation of the physical distribution system, where it models the distribution network and the control logic of system controllers as they are. In other words, the power flow solver model acts as a special machine learning model that has a full mathematical

representation of the true physical distribution system. Although the power flow solver is a highly accurate model, it contains too much details of the physical system which makes the algorithm too complicated and time consuming to solve.

The general machine learning algorithms are often treated as black boxes that can be applied to many distinct problems without knowing the mechanism of the true physical systems. Compared with the power flow solver which contains the full knowledge of the distribution system, the general machine learning algorithms, such as regression and clustering algorithms, contain too little or no knowledge of the physical system. This explains why they experience various challenges in creating an efficient black box that can follow the highly discrete power system behaviors in QSTS simulation.

As a result, the key for developing a fast QSTS simulation solution is developing a machine learning model that contains the right amount of the physical system knowledge. On the one hand, including too much physical system knowledge in the model will slow down the simulation. On the other hand, including too little physical system knowledge in the model will reduce the capability of the model to follow true system behaviors.

In this section, we propose a plane-based machine learning model by incorporating the right amount of human knowledge of the electrical distribution system. The new model contains the right amount of physical system knowledge that improves the computational speed and accuracy simultaneously. Two pieces of knowledge are used in the proposed algorithm: voltage sensitivity and system controllers' control logic. The voltage sensitivity will allow us improve the model computational speed. The system controllers' control logic will allow the model to estimate the system state transitions without solving power flow.

The plane-based machine learning model gets the name because the model can be visualized using the previous graphic model, where the power flow solutions can be represented by a series of planes. The proposed machine learning model can safely bypass the time-consuming process of solving power flows to speed up the QSTS simulation.

5.4.1 Sensitivity Model of System Controllable Elements

Both the regulator and capacitor controls are driven by bus voltages. We propose a linear approximation between bus voltages and system load. This approximation could be used to predict the behaviors of system controllers without solving the power flow.

Per discussion in Chapter 5.3.1, most system controllers' control logics depend on the system current, which is not linearly correlated with system loads due to the nonlinearity of the distribution network. However, in most distribution systems, we can take a linear approximation between the system current and the energy consumption. The linearized assumption is supported by reference [80], where a linearized assumption between bus voltage and system load is proposed. The authors also provide a tight upper bound of the linearization error by mathematical derivation.

5.4.2 Sensitivity Model for Multiple Load Profiles

In Chapter 5.3.1, a graphic model is used to represent the control logic of a power system regulator. In this chapter, we expand the one-load-profile cases into multiple-load-profile cases.

In hosting capacity analysis, the PV output profile usually has a much larger impact on distribution system controllers. Thus, we need to incorporate the PV output profile in

the graphic model. In the graphic model, we only need to add another dimension to the existing 2-D model in Figure 37. In Figure 45, the grey plane represents the scenarios of zero PV output similar to Figure 37. If we treat the PV output as a negative load, we can expect the same linear relationship between PV output and bus voltages, based on the linearization assumption. For a given regulator tap position, this will result in a plane-shaped representation between system load profiles and bus voltages. For example, plane $ABCD$ stands for the load-voltage correlation under a given regulator tap position, where line EH stands for zero PV output and line FG stands for maximum PV output. Like the 2-D case, the system controller setting forces the input voltage $V_{regCtrl}$ within the voltage band (V_{regMin}, V_{regMax}). Thus, the plane $ABCD$ is further curtailed by the two yellow planes representing the voltage limits of system controllers, which results in a contained plane $EFGH$.

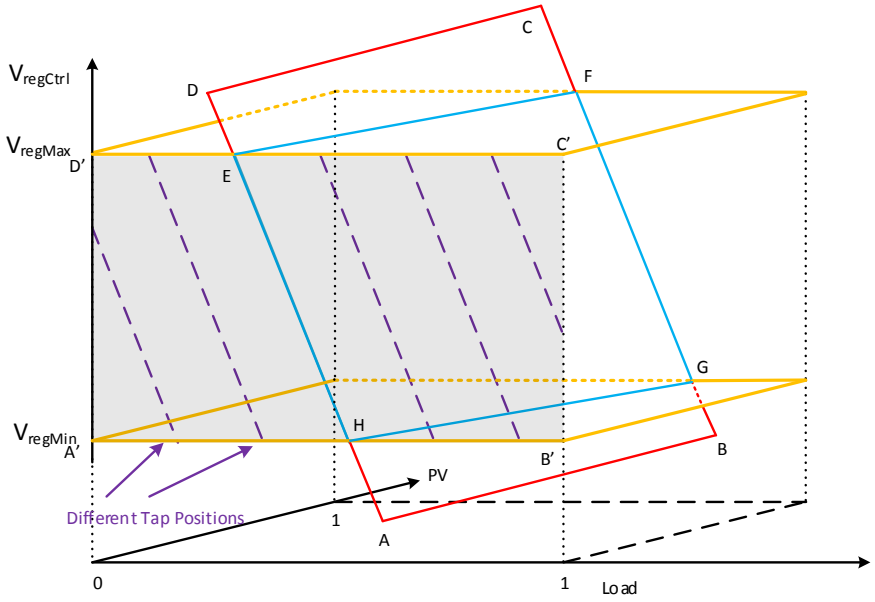

Figure 45. System regulator model with two load profiles.

Figure 45 shows how a contained plane can represent the mapping between system energy consumption and voltage for a given regulator tap position. We can easily expand this further to all tap positions with a series of parallel planes each representing a distinct tap position as shown in Figure 46. Since both load and PV profiles range from zero to one, all the parallel planes are constrained in a grey cuboid.

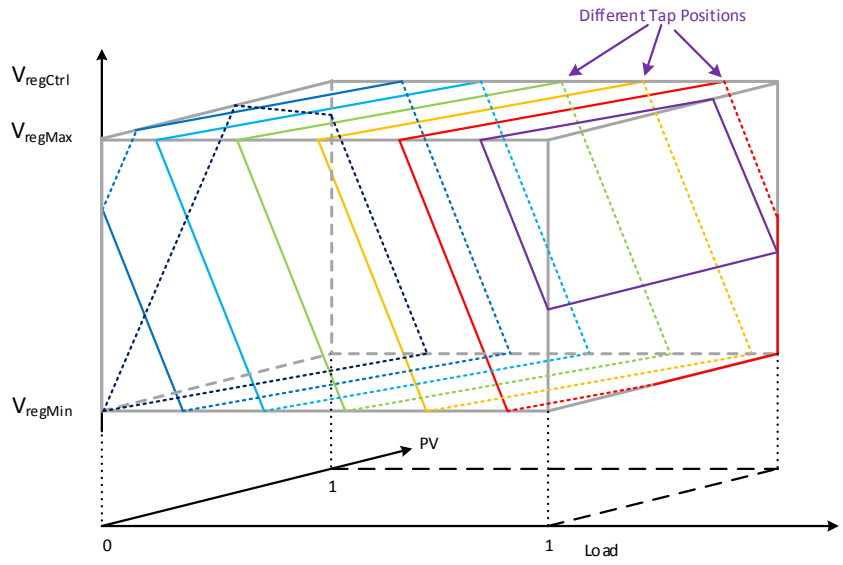

Figure 46. Multiple-plane model for different regulator tap positions.

Equipped with the graphic model shown in Figure 46, we can derive the voltage whenever the system controller state and load/PV values are given. In other words, the plane-based model can be used as an approximated power flow solver, where voltage can be estimated in no time. This can be shown in Figure 47. When the regulator tap position is given, we can identify a unique plane as $EFGH$. If the load and PV output are known as $Load_1$ and PV_1, then the voltage V_1 can be derived instantly on $EFGH$.

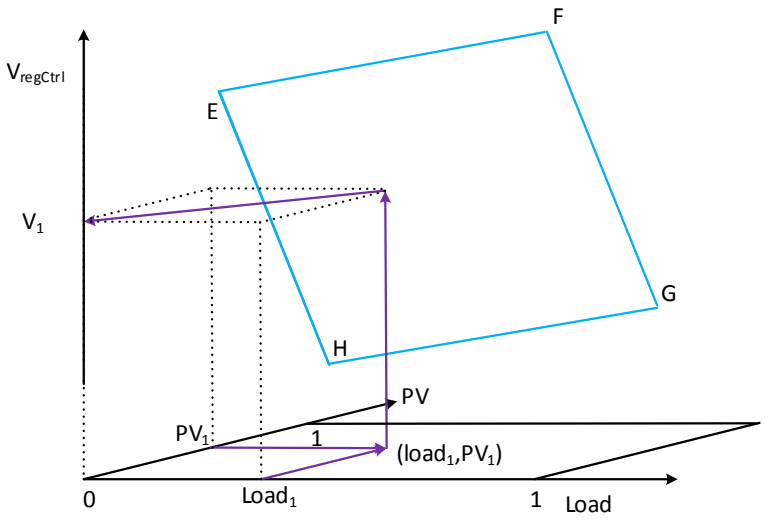

Figure 47. Using the graphic model to bypass solving power flow.

Like a regulator, a capacitor maintains the system voltage by switching the capacitor banks on and off based on the voltage at the regulated bus. When the capacitor is on and the voltage rises above the switch-off threshold V_{capOff}, the capacitor will switch off; when the capacitor is off and the voltage falls below the switch-on threshold V_{capOn}, the capacitor will switch on. Compared with regulators, capacitors only have two states: on and off. Similar graphical model applies to capacitors, as shown in Figure 48. The red plane represents the operational plane when the capacitor is off, and the blue plane where the capacitor is on. One decision boundary for the capacitor to switch on can be derived by the intersection of the plane $ABCD$ and V_{capOn}. Similarly, the other decision boundary for the capacitor to switch off can be derived by the intersection of the plane $EFGH$ and V_{capOff}. If we project the plane $ABCD$ and $EFGH$ down to the Load-PV space, we will get the decision boundaries of the two capacitor states as $A'B'C'D'$ and $E'F'G'H'$.

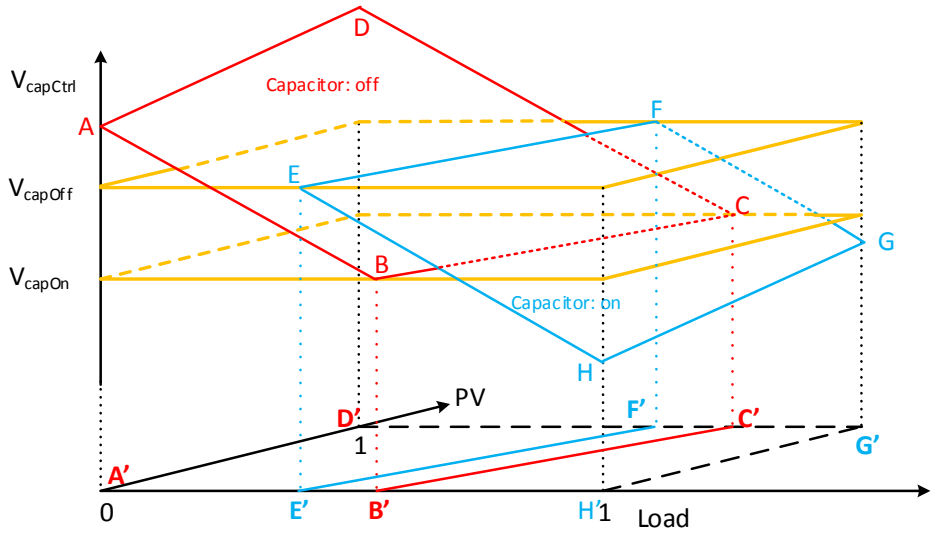

Figure 48. Graphic representation for capacitor controls.

In the previous discussion, we increase the number of load profiles from one to two by adding another dimension to the graphic model. However, the same technique applies to system with 3 or more load profiles. For example, the discrete linear model for a regulator with n load profiles can be mathematically presented as equation (27).

$$V_{regCtrl,i} = \boldsymbol{\beta}_i \boldsymbol{U} \qquad i = 1,2,\dots,m \tag{27}$$

where $V_{regCtrl,i}$ stands for the regulator control input voltage at tap position i; m stands for the total number of tap positions; $\boldsymbol{U}$ is a $(n+1) \times 1$ vector consisting of all load profiles. For example, in Figure 37, $\boldsymbol{U}^T = [load, 1]$; similarly, in Figure 46, $\boldsymbol{U}^T = [load, PV, 1]$. $\boldsymbol{\beta}_i$ is a $1 \times (n+1)$ vector standing for the coefficients of the linear model.

As long as the linear voltage sensitivity assumption holds, the proposed method does not have limitation on the number of load profiles in the system. A hyper plane can be built to accommodate any number of load profiles. This property is very appealing, because this

allows the QSTS simulation to incorporate multiple PV output profiles and even smart meter measurements.

5.4.3 System Events Prediction with Plane-based Model

The proposed plane-based model relies on a linear voltage-load correlation assumption and the graphic representation of the system controllers' control logic. In this section, we show how a plane-based model can help to predict the QSTS simulation system events. The most important justification of the long time high resolution QSTS simulation is predicting the system state transitions such as regulator and capacitor actions. With the proposed model, knowing system bus voltages is no longer necessary for the prediction of system state transitions through time.

According to the proposed graphic model, for a given regulator tap position, the correlation between $V_{regCtrl}$ and load/PV can be represented as a linearized plane in Figure 49. Line AB corresponds to $V_{regCtrl} = V_{regMin}$ and line DC corresponds to $V_{regCtrl} = V_{regMax}$. If we project the blue plane $ABCD$ down to the load-PV space, we get a red parallelogram $A'B'C'D'$. $A'B'C'D'$ is also the decision boundary of the current regulator tap position. For example, if the load and PV combination moves to the right of the red parallelogram, then we have $V_{regCtrl} < V_{regMin}$, which will cause a regulator tap switch-up action. Similarly, if the load and PV combination moves to the left of the red parallelogram, the regulator tap switch-down action will be triggered. In other words, if we get the decision boundary of a tap position on the load-PV plane, we no longer need to solve $V_{regCtrl}$ to predict the tap switch action. Instead, we only need to check whether the load and PV values are located within the red decision boundary.

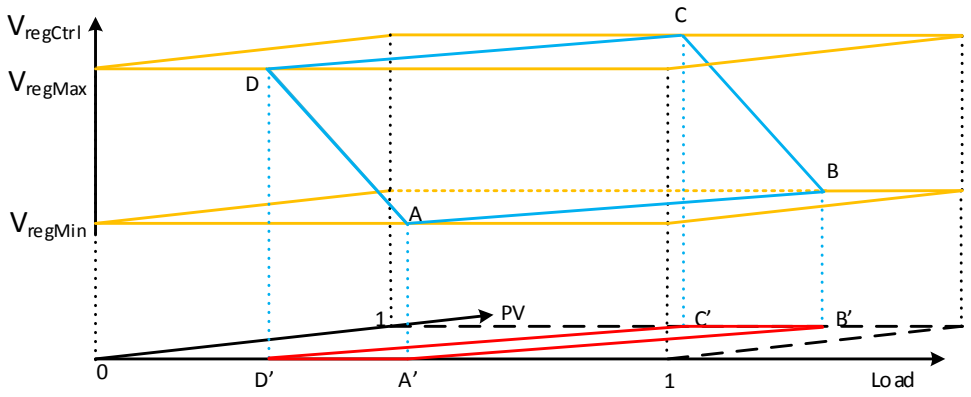

Figure 49. Decision boundary for a given regulator tap position.

In Figure 50, we ignore the voltage axis in Figure 46 and project the 3D plane into the Load-PV 2D space. We may notice that the planes are overlapping when projected down to the 2D space. For example, the plane $ABCD$ is overlapping with plane $A'B'C'D'$, which agrees with the multiple valid solution challenge discussed in the previous sections.

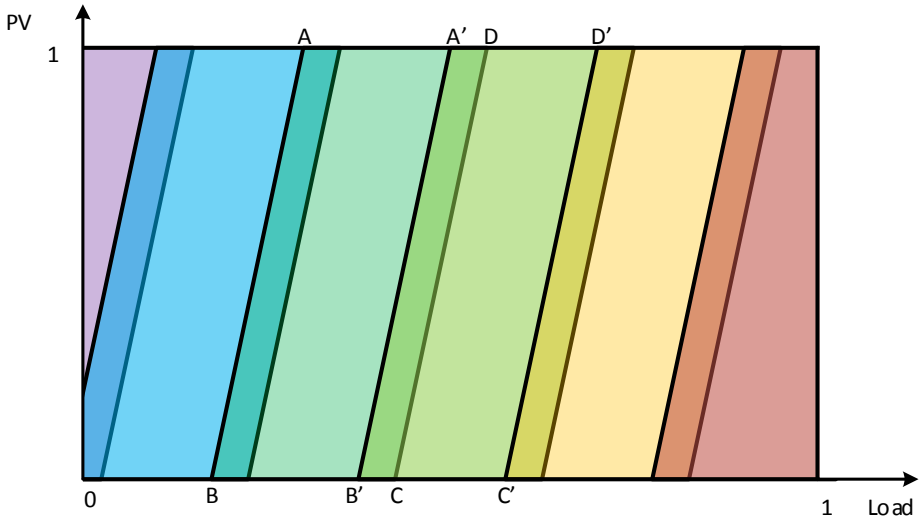

Figure 50. Reducing the dimensionality by ignoring the voltage dimension.

Figure 51 shows two adjacent decision boundaries: $A'B'C'D'$ and $E'F'G'H'$. To further illustrate how to use decision boundaries to predict system events, let us assume the

combination of load and PV values moves through the trajectory of $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e$ through time. The load and PV inputs start at point a, and the regulator tap was on the red position. The regulator stays put until load/PV moves to point b when the $V_{regCtrl}$ is equal to V_{regMin}. Since the load continues to drop after point b, $V_{regCtrl}$ becomes smaller than V_{regMin}, and a tap switch action is triggered, which boosts the $V_{regCtrl}$ to be above V_{regMin}, and the system now runs on the adjacent green plane. Similarly, when the load moves from c to e, $V_{regCtrl}$ becomes greater than V_{regMax} after point d. This will trigger a tap switch action at point d, and the system jumps from the green plane back onto the red plane.

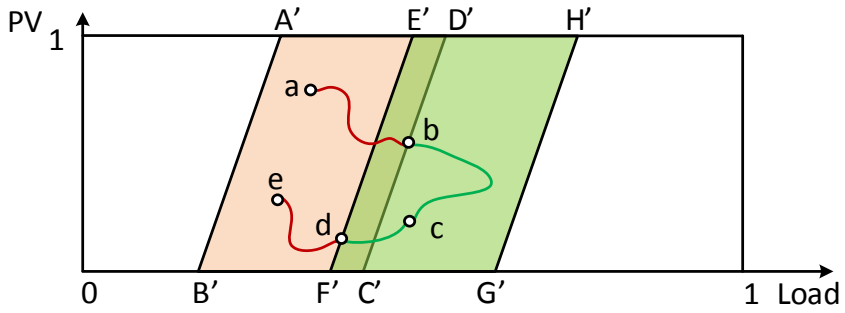

Figure 51. Predicting system events through decision boundaries.

In most distribution networks, multiple system controllers are presented. Due to the correlation among different system controllers, any action of a controller will have impacts on all other controllers. In this case, the proposed graphic model still applies and can be built for all system controllers. In fact, we just need to update the plane model whenever a controller takes action.

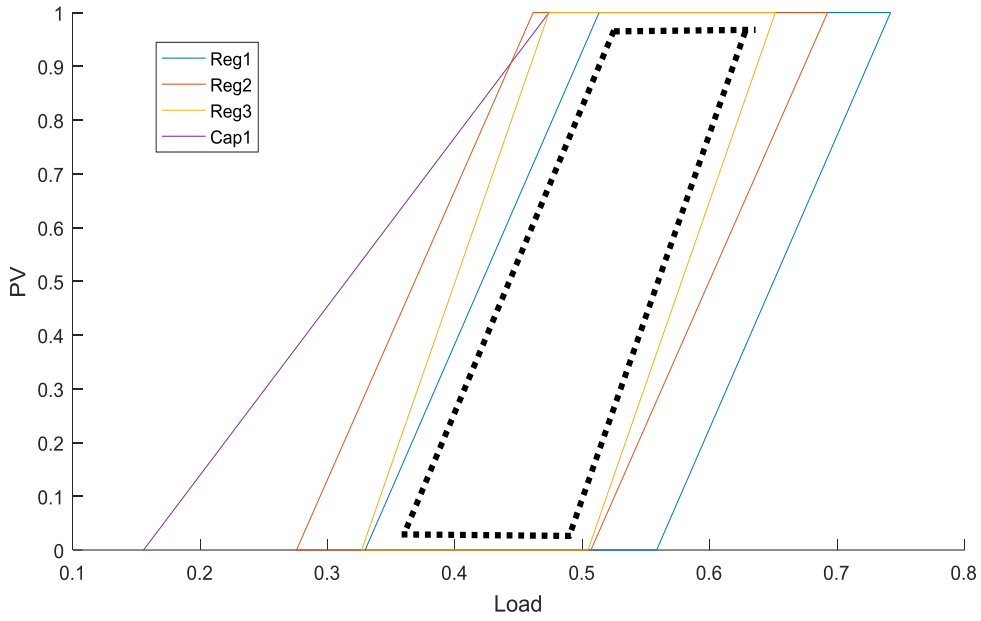

Figure 52. Decision boundaries for multiple system controllers.

Let us assume a distribution network has three regulators and one capacitor. We first build up the plane models for all controllers. Then, we combine the decision boundaries of all models as shown in Figure 52. The final decision boundary for the specific system state is the common area of all decision boundaries, shown as the black dashed lines. If the combination of load and PV moves out of the decision boundary, a system controller action will be triggered, and the system will move to another state with new a decision boundary.

5.4.4 Plane-based Model Parameter Estimation

The key for estimating the proposed sensitivity model is the estimation of the red decision boundary $A'B'C'D'$ or equivalently the blue plane $ABCD$, as shown in Figure 49. Line AD and line BC are determined by the PV output range, which is zero to one. Line AB and line CD are derived by the regulator setting V_{regMin} and V_{regMax}, which is known.

As a result, as long as the function of the plane $ABCD$ is known, the decision boundary can be drawn.

To uniquely determine a plane, mathematically, we only need three points, which is equivalent of solving three distinct power flows under the given system controller state. However, in practice, bus voltage and system load are not strictly linearly correlated. To increase the accuracy of the estimated plane, we use four distinct power flow solutions instead of three, to estimate the plane. Moreover, an iterative approach is developed to make sure the four power flows are solved near the edges of the decision boundary. This will minimize the error caused by the linearization approximation.

The iterative method keeps updating the four power flow solution locations to make sure the estimated plane is at least accurate at the edges of the decision boundary. Figure 53 shows the flow chart of the iterative method, where a certain iteration number is used as a stopping criteria. In Figure 53, we assume the system has n distinct load profiles and we are estimating the decision boundary of a single-phase regulator.

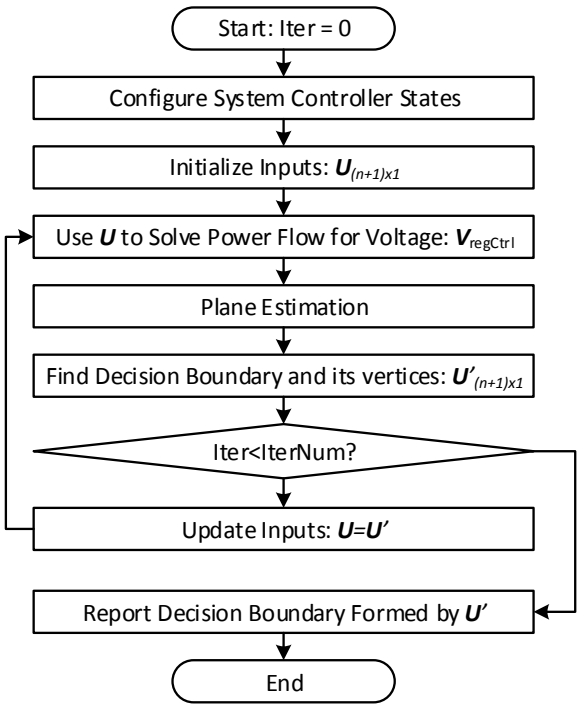

Figure 53. Flow chart of the iterative method.

Further illustration is shown in Figure 54, where we estimate the decision boundary of a regulator by two iterations. For the initial iteration, we pick four points with two random load levels combined with two scenarios where the PV output is 0 and 1. After solving the four power flows, we get four points A1, B1, C1, and D1, where a plane can be derived. The boundaries of the plane are calculated as A2-B2-C2-D2. On the second iteration, we use the load and PV value at A2, B2, C2, and D2 to calculate and update the plane and its boundaries. On the second iteration, the updated plane boundary A3-B3-C3-D3 is drawn using the plane estimated by power flow solutions solved at A2, B2, C2, and D2.

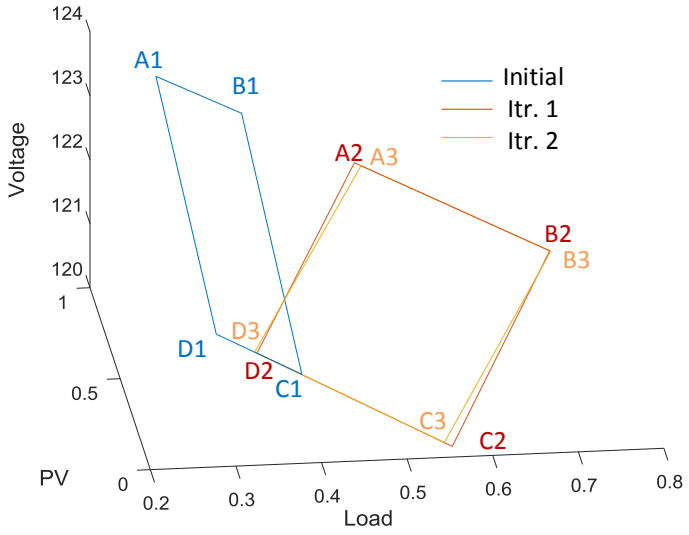

Figure 54. Iterative method for decision boundary accuracy improvement.

5.4.5 Plane-based Machine Learning Model for Fast QSTS Simulation

In this section, we piece the previous building blocks together and provide the whole flow chart of the proposed sensitivity model for fast QSTS simulation. As shown in Figure 55, the method starts with model initialization where the circuit is compiled. We store the computed plane models in a look-up table. Let $SimLength$ stands for the total simulation length. The only building block that requires solving power flow is the green portion of the flow chart, where a system event occurs and the plane model of the new system state has not been solved before. Often times, no power flow solve is involved if the simulation stays in the blue block where no system event occurs or the system state transits to a previously computed plane model.

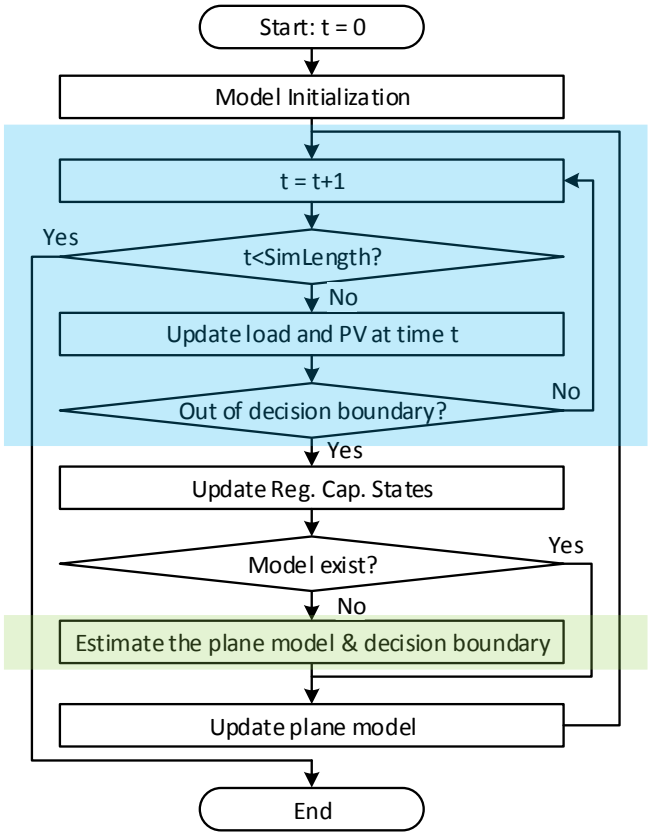

Figure 55. Flow chart for machine-aided fast QSTS simulation.

5.5 Test Results Analysis

The proposed machine aided fast QSTS simulation algorithm is tested on an IEEE 14 bus system. The test system has one load profile and one PV profile. Both profiles are of second-level resolution and collected in the field with distribution PMUs. Figure 56 shows a sample load and PV output profiles for 8 days. The test system is a modified IEEE 14-bus system, which has three independent single-phase regulators at the substation and one capacitor at the end of the feeder where a PV system is installed, as shown in Figure 57. The PV penetration of the network is set as 40 percent.

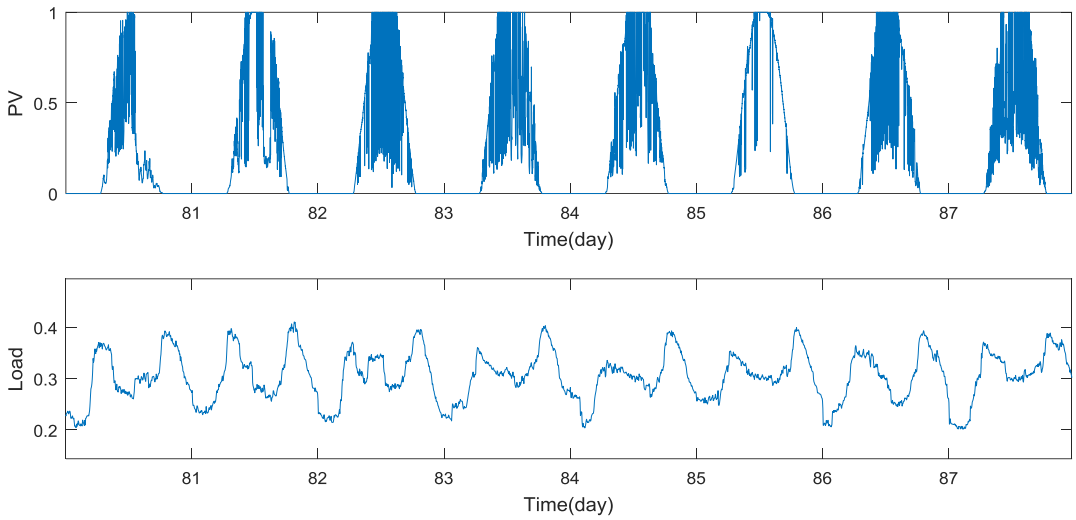

Figure 56. Sample PV output and load profiles.

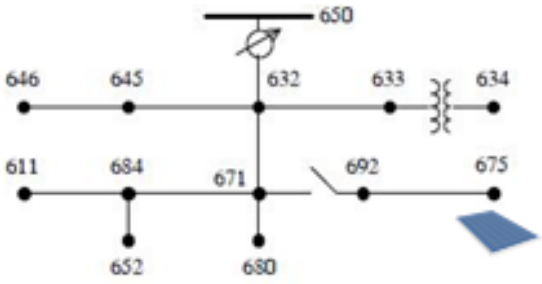

Figure 57. IEEE 14 bus system with a large PV system installed on bus 675.

To acquire the baseline simulation result, we run a yearlong 1-second QSTS simulation using the brute force method. For the 14-bus small system, the brute force simulation takes 13 minutes and 27 seconds. Figure 58 shows single-phase voltages at bus 675 in per unit with respect to load and PV profiles. Each dot represents a power flow solution for a specific time instance t of the QSTS simulation. We color the dots based on different regulator tap positions. All the dots associated with each tap position lie on separate surfaces which verified our graphic model. Since all these surfaces are flat, combined with the previous linear assumption, we refer to them as "planes". As the PV

and load change in the system, controller state changes, and the power flow solution will "jump" from one plane to another.

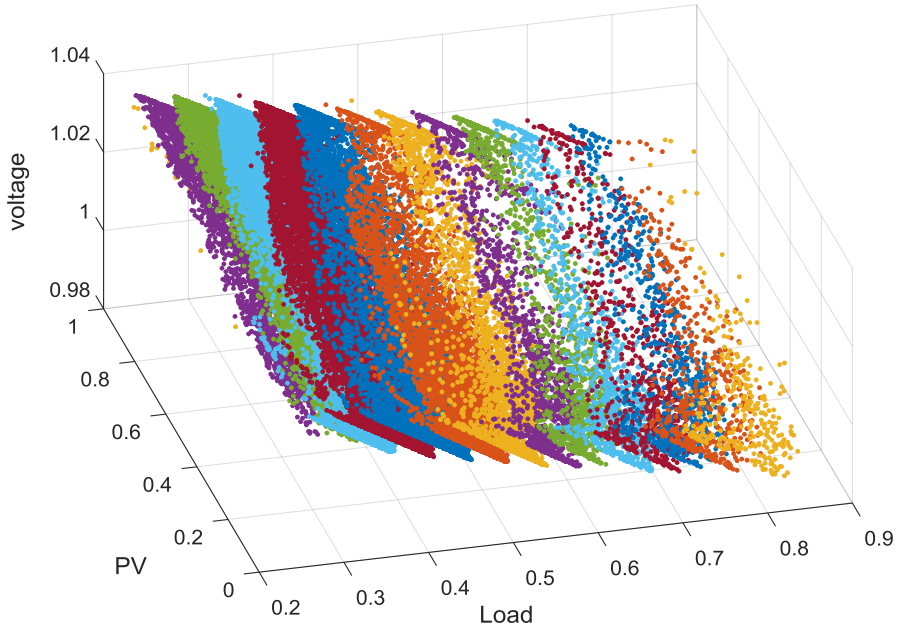

Figure 58. Bus 675 voltages for over 31 million power flows in QSTS simulation.

To test the accuracy of the proposed method, we run the same yearlong 1-second resolution QSTS simulation using the proposed model, and compare the simulation results with the brute force results. Figure 59 shows the system controllers' states from both the brute force method and the proposed method for 90 days. Since the controller states for all three regulators and one capacitor are overlapping for the two methods, the proposed method serves the purpose of predicting system state transitions very well. In hosting capacity analysis, the key index that high resolution QSTS can provide is the annual number of system state transitions or the regulator and capacitor actions. As the PV size continues to increase, it is necessary to run QSTS to make sure the plugged-in PV system

will not lead to system controller oscillations. The accuracy on annual system controller action number is shown in Table 7.

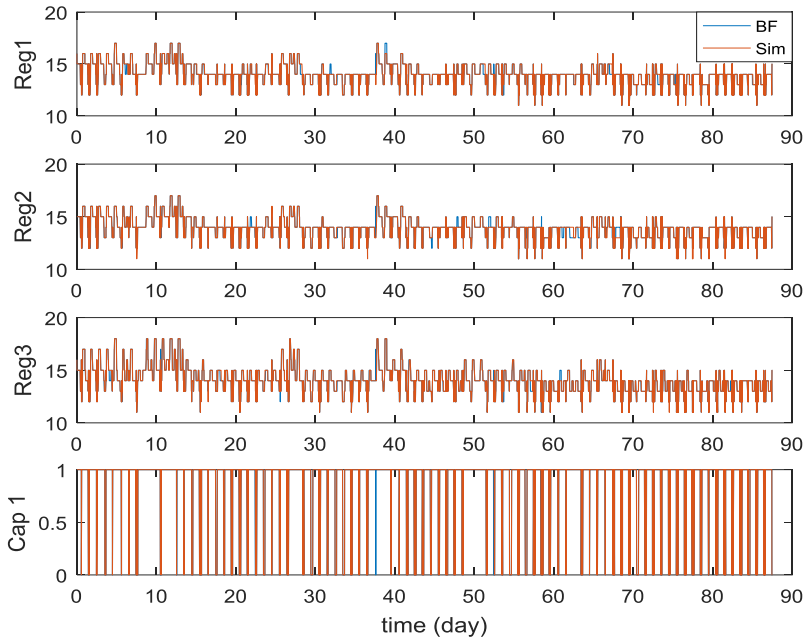

Figure 59. System controller state comparison between the proposed method and brute force method.

Table 7 illustrates how the iterative method proposed in 5.4.4 helps to improve the simulation accuracy. If the iterative method is not adopted, the model error is rather large but the computational time reduction is also significant. When we increase the number of iterations in estimating the decision boundary of the plane model, the simulation error decreases but the computational time increases slightly. Moreover, the simulation accuracy stabilizes after two iterations. This is because the estimated decision boundary converges to the true decision boundary rapidly with roughly two iterations. A more illustrative figure is shown in Figure 60 to demonstrate the trade-off between model accuracy and efficiency.

Table 7 – Model Accuracy and Efficiency Trade-off.

Num. of Iterations	Reg. Avg. Err (%)	Cap. Avg. Err (%)	Comp. Time (sec)	Comp. Time Reduction (%)
0	3.22	2.35	6.34	99.21
1	2.24	-5.19	6.47	99.20
2	1.91	-4.94	6.57	99.19
3	1.91	-4.94	6.75	99.16
4	1.91	-4.94	6.96	99.14
5	1.91	-4.94	7.09	99.12

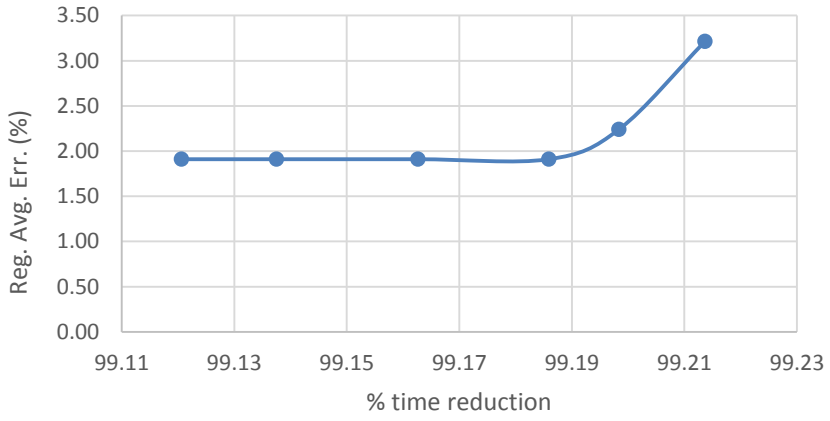

Figure 60. Model accuracy and computational time trade-off.

The current event-based algorithm works well on estimating system controller status through time. However, the annual over/under-voltage durations are also very important indices in evaluating feeder hosting capacity. The estimated annual over-voltage duration is 22.09 hours, which is very close to the brute force result of 22.13 hours; and the estimated under-voltage duration is 19.92 hours, while the brute force result is 11.47 hours. The estimation shows that the proposed algorithm can propose an approximate estimation on the duration of system bus voltage violation.

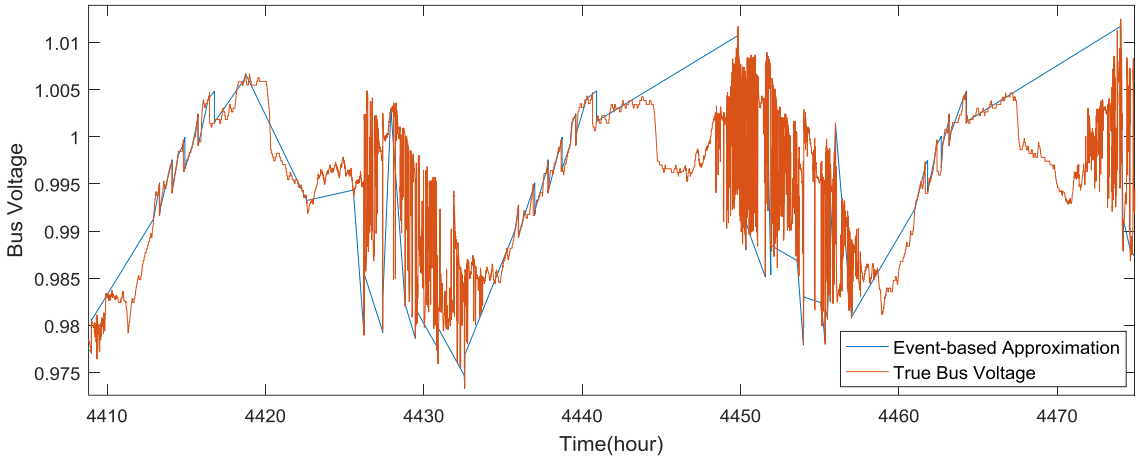

Figure 61. Event-based approximation for bus voltage.

5.6 Conclusion

In this Chapter, we first introduce QSTS simulation as the state-of-the-art hosting capacity analysis method, which provides a comprehensive and thorough evaluation of possible PV interconnection impacts. The major barrier that prevents the massive adoption of QSTS simulation is prohibitively large computational time. To speed up QSTS simulation, we first try out some general machine learning algorithms and analyze why conventional machine learning algorithms cannot meet our goals. Then, a machine-aided method based on a graphical representation of the distribution system is proposed. The new model feeds the machine with the right amount of human knowledge of the physical distribution system. The proposed method is validated through a test system with both high simulation speed and accuracy.

CHAPTER 6. CONCLUSION AND CONTRIBUTIONS

6.1 Contributions and Conclusion

Big data analytics and machine learning techniques are key enablers for enhanced smart grid operations and planning. The most important force to push power system analytics toward a data savvy endeavor is the ubiquitous data collected by tens of millions of newly installed sensors in the grid. The more data we collect, the more insights we can discover. This research acknowledges this transformation by presenting four distinct examples on how data analytics and machine learning algorithms serve to solve practical smart grid problems.

In the first example, a data-driven solution is provided to detect and estimate unauthorized residential PV systems. The method mines the consumer smart meter data for abnormal energy consumption behaviors. We use change-point detection in this application, pointing out also possible applications to other critical customer behaviors, such as energy theft, new EV adoption, and energy efficiency improvement. The method utilizes the concept of data fusion, where the combination of energy data and weather data provides additional confidence in estimating the sizes of unfiled residential PV systems.

In the second example, a stochastic model is developed to describe residential EV charging demand. This is an example of using smart meter historical data to enhance our understanding of customer energy consumption behaviors. Based on the long run simulation results of the stochastic model, we can justify that although the number of EVs continues to grow, the possibility of all EVs in a neighborhood charging simultaneously is

small. In fact, based on the proposed model, it is very unlikely to have more than 25% of the total residential EVs charging at the same time.

In the third example, we introduce a time-variant load model, which captures the changes of electrical load properties through time. Unlike the traditional static load model in which the load's P-V and Q-V properties are function of voltage, the proposed model writes the load's P-V and Q-V properties as a function of voltage and time. This is because the time-variant model takes advantages of the time information of the smart meter measurements.

In the last example, we develop a machine-aided hosting capacity analysis method based on fast QSTS simulation. In this example, we show that pre-existing and general machine learning algorithms might not always work for our specific smart grid problems. It is necessary to input knowledge of the physical smart grid to the machine to create a model that can achieve better performance for solving specific smart grid problems.

Finally, as more and more sensors are deployed in power energy systems, the value of big data analytics and machine learning is only going to increase. At the same time, future research efforts will continue to push our understanding of the grid forward as both the quality and the quantity of the data continue to improve and grow.

6.2 Future Work

In this dissertation, I have explored how big data analytics and machine learning can benefit power system operations and planning. There are still vast research opportunities remaining on this topic. Potential future work can be categorized into three categories.

First, data analytics will play a major role in demand response and tariff design. Although current demand response programs and tariff designs are based on some energy consumption data, they are far from intelligent and rely heavily on human subjective judgement. There is room for improvement, where the massive smart meter data can assist the understanding of energy consumption down to every single customer. For example, there is still no reliable algorithm to identify and estimate the critical load components such as HVAC, electrical vehicles, and PV systems. Future research could focus on analyzing energy meter measurements to gain a better understanding of customer energy consumption patterns and how likely they are to respond to incentives.

Second, distribution model parameter estimation is another major direction that can use a boost from the sensor measurements through data analytics. In practice, it is still common that utilities and system operators do not have access to accurate and detailed distribution network models. However, without the proper understanding of the network model as a foundation, more advanced applications such as hosting capacity analysis or demand response cannot be implemented safely. Future research could focus on further enhancement of the calibration of distribution network models, including system topology and secondary element parameters.

Third, asset management will benefit from the advanced machine learning algorithms and data analytics. Power system asset management is a very pressing issue especially when facing an aging network. Data analytics can effectively detect asset failure risks through predictive analysis, thus improving network reliability, maximizing asset utilization, and optimizing network availability. Future research could focus on asset health estimation and system event prediction using the real-time measurements collected by

power system sensors. These research efforts will facilitate system operators in making the right decision before any potential system failure occurs.

REFERENCES

[1] "Clean Power Plan for Existing Power Plants," U.S. Energy Information Administration, Retrieved August 3, 2015.

[2] "Annual Energy Outlook 2016, with Projections to 2040," U.S. Energy Information Administration, Chapter 5. Electricity, August 2016.

[3] "International Energy Outlook 2016," U.S. Energy Information Administration, Chapter 5. Electricity, May 11, 2016.

[4] "Global Plug-in Light Vehicles Sales Increased By About 80% in 2015," Argonne National Laboratory, United States Department of Energy, March 28, 2016.

[5] "Algorithm for Screening Phasor Measurement Unit Data for Power System Events and Categories and Common Characteristics for Events Seen in Phasor Measurement Unit Relative Phase-Angle Differences and Frequency Signals," National Renewable Energy Laboratory, August 2013.

[6] "Utility-Scale Smart Meter Deployments: A Foundation for Expanded Grid Benefits," Innovation Electricity Efficiency Institute, the Edition Foundation, August 2013.

[7] Langley, P. and Simon, H.A. "Applications of machine learning and rule induction." Communications of the ACM 38, no. 11 (1995): 54-64.

[8] Y Singh, PK Bhatia, O Sangwan, A review of studies on machine learning techniques, International Journal of Computer Science and Security, 2007

[9] O. A. S. Youssef, "Combined fuzzy-logic wavelet-based fault classification technique for power system relaying," Power Delivery, IEEE Trans. on, vol. 19, no. 2, pp. 582-589, Apr 2004.

[10] P. Kind, "Disruptive Challenges: Financial and Strategic Responses to a Changing Retail Electric Business," The Edison Electric Institute (EEI), Jan. 2013.

[11] P. Fairley, "Hawaii's Solar Push Strains the Grid," Available online: http://www.technologyreview.com/ news/534266/hawaiis-solar-push-strains-the-grid/, Jan. 20, 2015.

[12] "HECO customers asked to disconnect unauthorized PV systems," Available online: http://khon2.com/ 2014/09/05/heco-customers-asked-to-disconnect-unauthorized-pv-systems/, Sept. 5, 2014.

[13] Transient Over-Voltage Mitigation: Explanation and Mitigation Options for Inverter-Based Distributed Generation Projects, Hawaii Electric Companies, Feb 24, 2014.

[14] California Solar Permitting Guidebook, Solar Permitting Work Group, The Governor's Office of Planning and Research, June 2012.

[15] Guide to Renewable Energy Facility Permits in the state of Hawaii, the Hawaii Clean Energy Initiative (HCEI), Apr. 2015.

[16] 2014 Smart grid system report, U.S. Department of Energy (DOE), Available online: http://energy.gov/oe/downloads/2014-smart-grid-system-report-august-2014, Aug. 2014.

[17] In the Matter of Arizona Public Service Company's Application for Approval of Net Metering Cost Shift Solution, ACC Docket – Arizona Corporation Commission Docket No. E-01345A-13-0248. Dec. 8 2013.

[18] PHOTON, "Flanders: Illegally installed PV systems could outnumber systems installed under GC scheme", Available online at: http://www.photon.info/photon_news_detail_en.photon?id=78701, July 19, 2013.

[19] P. Denholm, and R. Margolis, "Supply Curves for Rooftop Solar PV-Generated Electricity for the United States," National Renewable Energy Laboratory (NREL), Nov. 2008.

[20] X. Zhang, Z. Bie, and G. Li, "Reliability Assessment of Distribution Networks with Distributed Generations using Monte Carlo Method." Energy Procedia, vol. 12, pp. 278-286, 2011.

[21] Baran, M.E.; Hooshyar, H.; Zhan Shen; Huang, A., "Accommodating High PV Penetration on Distribution Feeders," in Smart Grid, IEEE Transactions on , vol.3, no.2, pp.1039-1046, June 2012

[22] M.J. Reno, K. Coogan, R.J. Broderick, J. Seuss, and S. Grijalva, "Impact of PV Variability and Ramping Events on Distribution Voltage Regulation Equipment," in IEEE Photovoltaic Specialists Conference, 2014.

[23] Peng Li; Xiaomeng Yu; Jing Zhang; Ziheng Yin, "The H ∞ Control Method of Grid-Tied Photovoltaic Generation," in Smart Grid, IEEE Transactions on , vol.6, no.4, pp.1670-1677, July 2015

[24] Samadi, A.; Soder, L.; Shayesteh, E.; Eriksson, R., "Static Equivalent of Distribution Grids With High Penetration of PV Systems," in Smart Grid, IEEE Transactions on , vol.6, no.4, pp.1763-1774, July 2015

[25] Ravindra, H.; Faruque, M.O.; McLaren, P.; Schoder, K.; Steurer, M.; Meeker, R., "Impact of PV on distribution protection system," in North American Power Symposium (NAPS), 2012 , vol., no., pp.1-6, 9-11 Sept. 2012

[26] X. Zhang and S. Grijalva, "A Data-Driven Approach for Detection and Estimation of Residential PV Installations," in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2477-2485, Sept. 2016.

[27] Hong, T, "Energy Forecasting: Past, Present and Future," Foresight: The International Journal of Applied Forecasting, vol. 32, pp. 43-48, 2014.

[28] X. Zhang, S. Grijalva, and J.R. Matthew, "A time-variant load model based on smart meter data mining." in Proc. 2014 IEEE Power and Energy Society General Meeting.

[29] X. Zhang, and S. Grijalva, "An Advanced Data Driven Model for Residential Electric Vehicle Charging Demand," in Proc. 2015 IEEE Power and Energy Society General Meeting.

[30] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," Neural Networks, vol. 43, pp. 72-83, 2013.

[31] Huschke, Ralph E. (1970) [1959]. "Cloud cover". Glossary of Meteorology (2nd ed.). Boston: American Meteorological Society. Retrieved August 24, 2013.

[32] S. M. Ali, and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," Journal of the Royal Statistical Society, Series B, vol. 28, no. 1, pp. 131–142, 1966.

[33] R. Lund, L. Wang, Q. Lu, J. Reeves, C. Gallagher, Y. Feng, "Changepoint detection in periodic and autocorrelated time series," J. Climate, vol. 20, pp. 5178–5190, 2007.

[34] P. Khandelwal, K.K. Singh, B.K. Singh, and A. Mehrotra, "Unsupervised Change Detection of Multispectral Images using Wavelet Fusion and Kohonen Clustering Network," International Journal of Engineering and Technology, 2013.

[35] J. Kyong, and I. Han, "An intelligent clustering forecasting system based on change-point detection and artificial neural networks: application to financial economics," System Sciences, Proc. of the 34th Annual Hawaii International Conference on, vol. 8, pp. 3-6, Jan. 2001.

[36] T. Kanamori, T. Suzuki, M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation." Machine Learning, vol. 86, no. 3, pp. 335-367, 2012.

[37] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, "Relative density-ratio estimation for robust distribution comparison," Neural Computation, vol. 25, no. 5, pp. 1324-1370, 2013.

[38] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," Proc. of the IEEE, vol. 90, no. 7, pp. 1151-1163, 2002.

[39] J. L. Rodgers, and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," The American Statistician, vol. 42, no. 1, pp. 59-66, 1988.

[40] Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003.

[41] M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research," Malawi Medical Journal : The Journal of Medical Association of Malawi, vol. 24, no. 3, pp. 69-71, 2012.

[42] J.I. Odiase, and S.M. Ogbonmwan, "Correlation analysis: exact permutation paradigm," МАТЕМАТИЧКИ ВЕСНИК, vol. 59, pp. 161-170, 2007.

[43] Resampling (statistics): Permutation test, Wikipedia, available online at: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests, Dec. 16, 2015.

[44] F.R. Martins, S.A.B. Silva, EB. Pereira, and S.L. Abreu, "The influence of cloud cover index on the accuracy of solar irradiance model estimates," Meteorology and Atmospheric Physics, vol. 3, no. 99, pp. 169-180, 2008.

[45] R. McGill, J. W. Tukey, and W.A. Larsen, "Variations of Boxplots," The American Statistician. vol. 32, no. 1, pp. 12–16, 1978.

[46] Bradley, T. H. and A. A. Frank, "Design, demonstrations and sustainability impact assessments for plug-in hybrid electric vehicles," Renewable and Sustainable Energy Reviews, vol. 13(1), pp. 115-128, 2009.

[47] Carley, S., et al., "Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cites," Transportation Research Part D: Transport and Environment, vol. 18(0), pp. 39-45, 2013.

[48] Jeff Cobb, "Global Plug-in Car Sales Now Over 600,000," HybridCars.com, Retrieved 2014-10-23.

[49] Clement-Nyns, K., et al., "The Impact of Charging Plug-In Hybrid Electric Vehicles on a Residential Distribution Grid," IEEE Trans. Power Systems, vol. 25(1), pp. 371-380, 2010.

[50] Sortomme, E., et al. "Coordinated Charging of Plug-In Hybrid Electric Vehicles to Minimize Distribution System Losses," IEEE Trans. Smart Grid, vol. 2(1), pp. 198-205, 2011.

[51] Qian, Kejun, et al., "Modeling of load demand due to EV battery charging in distribution systems," IEEE Trans. Power Systems, vol. 26(2), pp. 802-810, 2011.

[52] Gan, L. and Z. Xiao-Ping, "Modeling of Plug-in Hybrid Electric Vehicle Charging Demand in Probabilistic Power Flow Calculations," IEEE Trans. Smart Grid, vol. 3(1), pp. 492-499, 2012.

[53] Peng, H., et al., "A Novel Coordinative Resident Electric Vehicle Charging Mechanism," in 2012, Power and Energy Engineering Conf.

[54] Said, D., et al., "Queuing model for EVs charging at public supply stations," in 2013, Wireless Communications and Mobile Computing Conf., vol. 1(5), pp. 65-70

[55] Turitsyn, K., et al., "Robust Broadcast-Communication Control of Electric Vehicle Charging," in Proc. 2010, IEEE Smart Grid Communications International Conf., vol. 4(6), pp. 203-207.

[56] Zhang, X., et al., "A time-variant load model based on smart meter data mining," in Proc. 2014 IEEE Power Engineering Society General Meeting, vol. 1(5), pp. 27-31.

[57] Pecan Street Dataport - A Universe of Data, Available Around the World. Available: https://dataport.pecanstreet.org/.

[58] D. Anderson, "An Introduction to Management Science: Quantitative Approaches to Decision Making," 13nd ed., Cengage Learning, 2011, pp. 256.

[59] Feller, W., "Introduction to Probability Theory and Its Applications," vol. 2, 2nd ed., Wiley, 1971, Section 1.3.

[60] Yijia, C., et al., "An Optimized EV Charging Model Considering TOU Price and SOC Curve," IEEE Trans. Smart Grid vol. 3(1), pp. 388-393, 2012.

[61] Altiok, T., et al., "Simulation Modeling and Analysis with ARENA," Elsevier, 2007, pp. 161.

[62] IEEE Committee, Load representation for dynamic performance analysis. IEEE Transections on Power Systems, 1993, (2),472-482.

[63] IEEE Task Force Report, “Load Representation for Dynamic Performance Analysis,” Paper 92WM126-3 PWRD, presented at the IEEE PES Winter Meeting, New York, January 26-30, 1992.

[64] T. Frantz, T. Gentile, S. Ihara, N. Simons, M. Waldron, “Load Behaviour Observed in LILCO and RG&E Systems”, IEEE Trans., Vol. PAS-103, No. 4, April 1984.

[65] S.A.Y. Sabir, D.C Lee, “Dynamic Load Models Derived from data Acquired During System Transients,” IEEE Trans., Vol. PAS-101, September 1982, pp 3365 to 3372.

[66] Vaahedi, E., et al. (1987). "Load Models for Large-Scale Stability Studies from End-User Consumption." Power Systems, IEEE Transactions on 2(4): 864-870.

[67] Price, W. W., et al. (1988). "Load modeling for power flow and transient stability computer studies." Power Systems, IEEE Transactions on 3(1): 180-187.

[68] I. R. Navarro, “Dynamic load models for power systems,” Ph.D. dissertation, Dept. of Ind. Elect. Eng. and Auto., Lund University, Lund, 2002.

[69] Cover, T.M. and J.A. Thomas. "Elements of Information Theory," Wiley, 1991.

[70] Johnson, D.H. and S. Sinanovic. "Symmetrizing the Kullback-Leibler distance." IEEE Transactions on Information Theory

[71] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[72] Dingding Wang (Sch. of Comput. Sci., Florida Int. Univ., Miami, FL, USA); Ding, C.; Tao Li Source: Machine Learning and Knowledge Discovery in Databases. Proceedings European Conference, ECML PKDD 2009, p 506-21, 2009

[73] Bokhari, A.; Alkan, A.; Dogan, R.; Diaz-Aguilo, M.; de Leon, F.; Czarkowski, D.; Zabar, Z.; Birenbaum, L.; Noel, A.; Uosef, R.E., "Experimental Determination of the ZIP Coefficients for Modern Residential, Commercial, and Industrial Loads," Power Delivery, IEEE Transactions on , vol.PP, no.99, pp.1,1

[74] Kundur, (1994). Power System Stability And Control, McGraw-Hill Education (India) Pvt Limited.

[75] Solar Market Insight Report 2014, GTM Research/SEIA: U.S. Solar Market Insight.

[76] J. Peppanen, S. Grijalva, M. J. Reno and R. J. Broderick, "Secondary circuit model creation and validation with AMI and transformer measurements," 2016 North American Power Symposium (NAPS), Denver, CO, 2016, pp. 1-6.

[77] J. Peppanen, M. J. Reno, R. J. Broderick and S. Grijalva, "Distribution System Model Calibration with Big Data from AMI and PV Inverters," in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2497-2506, Sept. 2016.

[78] M. J. Reno and R. J. Broderick, "Predetermined time-step solver for rapid quasi-static time series (QSTS) of distribution systems," 2017 ISGT.

[79] M. J. Reno, J. Deboever, and B. Mather, "Motivation and requirements for quasi-static time series (QSTS) for distribution system analysis," 2017 PES GM.

[80] S. Volognani and S. Zampieri, "On the existence and linear approximation of the power flow solution in power distribution networks," IEEE Transactions on Power Systems 31.1 (2016): 163-172.